STATISTICAL ASSESSMENT OF MEDICATION ADHERENCE DATA: A TECHNIQUE TO ANALYZE THE J-SHAPED CURVE

by

Jeffrey M. Rohay

B.S. in Psychology, University of Pittsburgh, 1984

M.S.I.S. Information Science, University of Pittsburgh, 1988

M.S. Biostatistics, University of Pittsburgh, 1993

Submitted to the Graduate Faculty of

Biostatistics in

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jeffrey M. Rohay

It was defended on

November 18, 2009

and approved by

Dissertation Advisor: Gary M. Marsh, Ph.D., F.A.C.E., Professor and Interim Chair, Director, Center for Occupational Biostatistics and Epidemiology, Graduate School of Public Health, University of Pittsburgh

Stewart J. Anderson, Ph.D., Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Vincent C. Arena, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ada Youk, Ph.D., Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Jacqueline Dunbar-Jacob, Ph.D., R.N., F.A.A.N., Dean, School of Nursing, Professor of Nursing, Epidemiology, Psychology and Occupational Therapy, University of Pittsburgh

STATISTICAL ASSESSMENT OF MEDICATION ADHERENCE DATA: A TECHNIQUE TO ANALYZE THE J-SHAPED CURVE

Jeffrey M. Rohay, Ph.D.

University of Pittsburgh, 2009

ABSTRACT

Medication non-adherence impacts public health by impeding the evaluation of medication efficacy, decreasing improvement and/or increasing morbidity in patients, while increasing health care costs. As a result, intervention studies are designed to improve adherence rates. Medication adherence is J-shaped in nature with many people taking their medication completely, a significant proportion taking no medication, and a substantial proportion taking their medication on some intermittent schedule. Therefore, descriptive statistics and standard statistical techniques (e.g., parametric t-tests, non-parametric Wilcoxon Rank Sum tests, and dichotomization) can provide misleading results. This study developed and evaluated a method to more accurately assess interventions designed to improve adherence. Better evaluation could lead to identifying new interventions that decrease morbidity, mortality, and health care costs.

Parametric techniques utilizing a Gaussian distribution are inappropriate as J-shaped adherence distributions violate the normality assumption and transformations fail to induce normality. Additionally, measures of central tendency fail to provide an adequate depiction of the distribution. While non-parametric techniques overcome distributional problems, they fail to adequately describe the distribution's shape. Similarly, dichotomizing data results in a loss of information, making small improvements impossible to detect.

Using a mixture of beta distributions to describe adherence measures and the expectation-maximization algorithm, parameter and standard error estimates of this distribution were produced. This technique is advantageous as it allows one to both describe the shape of the distribution and compare parameter estimates. We assessed, via simulation studies, α-levels and power for this new method as compared to standard methods. Additionally, we applied the technique to data obtained from studies designed to increase medication adherence in rheumatoid arthritis patients.

Via simulations, the mixed beta model was shown to adequately depict adherence distributions. This technique performed better at distinguishing datasets, exhibiting power ranging from 66% to 92% across samples sizes. Additionally, α-levels for the new technique were reasonable, ranging from 3.4% to 5.4%. Finally, application to the "Adherence in Rheumatoid Arthritis: Nursing Interventions" studies produced parameters estimates and allowed for the comparison of interventions. The p-value for this new test was 0.0597, compared to 0.20 for the t-test.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express sincere and deep gratitude to my committee members, family, friends, and colleagues for support, assistance and encouragement while completing my dissertation. I would like to thank Dr. Marsh for his encouragement and patience as I completed this process. I would like to thank Dr. Anderson for taking the time to meet with me to help guide my research. I would also like to thank Dr. Arena and Dr. Youk for their excellent comments, guidance, and encouragement while I was working on this project.

I would especially like to thank Dr. Dunbar-Jacob for her assistance, guidance, support, and encouragement during this process, as well as throughout my entire career. I appreciate the opportunity to work on an exciting topic and hope to have much future collaboration.

To Dr. Rolf Jacob – I appreciate the support and encouragement you provided when I first entered the research field. I would not have completed this journey without your guidance, patience, and encouragement while taking my first steps in my research career.

I would like to thank my parents for always encouraging me to pursue my dreams and my brother for always supporting and encouraging me. I'd like to thank my son Matthew for being patient while I was working on this project.

Finally, I would like to thank my wife, Cindy. She was always there to cheer me up and encourage me through all of the ups and downs of completing this dissertation. I could not have done it without her.

1.0 INTRODUCTION

1.1 STATEMENT OF PROBLEM

Allport (1934) was among the earliest to describe the J-shaped distribution in his paper on telic continuums. A telic continuum can be defined as "one of the amount of fulfillment of the purpose of a common or prescribed act" (Solomon, 1939). In his paper, Allport noted that when the proportion of persons completing various portions of a behavior is plotted, the curve tends to follow a J-shape. Medication adherence measures are clearly of this class. Many people faithfully take their medication completely while a significant proportion of people fail to take any medication. A substantial proportion of people intermittently take their medication at a rate spread out between these two extremes. Medication non-adherence impacts public health in several manners. It has been cited as a major reason for the lack of improvement of patients on a prescribed drug regimen (Zygmunt, Olfson, Boyer, & Mechanic, 2002; Bartlett J. A., 2002; McDonald, Garg, & Haynes, 2002) as well as increased morbidity (Bartlett, Lukk, Butz, Lampros-Klein, & Rand, 2002). Additionally, accurate assessment of adherence is crucial in evaluating the efficacy of new treatments (Olivieri, Matsui, Hermann, & Koren, 1991). Finally, non-adherence has been estimated to increase health care costs by $290 billion (Dolan, 2009). As a result studies have been conducted that are designed to increase medication adherence

(Esposito, 1995; Dunbar-Jacob, Sereika, Burke, Starz, Rohay, & Kwoh, 1995; Bartlett, Lukk, Butz, Lampros-Klein, & Rand, 2002).

In these studies the adherence measure serves as the dependent variable and statistical analyses consist of parametric techniques such as t-test or ANOVA, non-parametric techniques such as Wilcoxon Rank-Sum test or Kolmorgov-Smirnov test, or χ^2 test of association on the dichotomized adherence measure. Each of these techniques has limitations. Parametric techniques utilizing a Gaussian distribution are not appropriate methods of analysis as the J-shaped distribution of adherence violates the assumption of normality and attempts to transform data fail to induce normality. Additionally, measures of central tendency fail to provide an adequate depiction of the distribution. While non-parametric techniques overcome distributional problems they still fail to adequately describe the shape of the distribution. Similarly, dichotomizing the data results in a loss of information, making small improvements in adherence impossible to detect.

Thus, an analytic technique is needed to overcome these limitations that will produce a proper model of the data and by providing parameter estimates and associated standard errors, allow for statistical comparisons of model. We propose using the following mixed beta distribution to model these data:

$$x_i = \pi Beta(1,\beta) + (1-\pi) Beta(\alpha,1)$$
$$where\ 0 \le \pi \le 1 is the mixing\ paramenter$$

Equation 1.1

The beta distribution is an excellent candidate for modeling adherence data as it is a continuous distribution on the interval 0 to 1. A beta distribution with parameters 1 and β (>1) produce a strictly decreasing curve representing the left hand portion of the distribution, while

beta distributions with parameters α (>1) and 1 produce a strictly increasing curve, representing the right hand portion of the curve (Cassella & Berger, 1990). The parameter π determines the proportion of data points assumed to be from each component of the distribution. It will be shown that this model produces a J-shaped distribution.

We will develop the Expectation-Maximization Algorithm for Adherence Parameters Estimation (EMAAPE) for estimating the parameters of the mixed beta distribution. The standard errors of the estimates will be computed using the Jackknife procedure. The advantage of this technique is that it allows one to both describe the shape of the distribution and to compare parameter estimates for two distributions to appropriately determine if two interventions differ statistically with respect to adherence.

Via simulation studies, α-levels and power for this new method will be compared to other standard methods. Additionally, bias will be assessed by comparing the parameter estimates to their known value. Finally, the proposed technique of analysis will be applied to an existing dataset comparing two interventions designed to increase medication adherence in the treatment of rheumatoid arthritis.

Specifically, in this dissertation we:

1. Generate data from a mixed beta distribution and demonstrate the resultant distribution is J-shaped.
2. Demonstrate that transformations fail to induce normality.
3. Develop the Expectation-Maximization Algorithm for Adherence Parameters Estimation (EMAAPE) to estimate the parameters and their standard errors.

4. Compare data from two distributions using established techniques such as t-tests, χ^2 tests, Wilcoxon Rank-Sum test, Kolmorgov-Smirnov test, Log-Rank tests, and Ordinal Logistic Regression.
5. Compare data from two distributions using parameter estimates from the EMAAPE.
6. Demonstrate the proposed technique is superior to established techniques in discerning differences between groups, particularly among smaller datasets (n=100-200) which would typically be observed in studies designed to increase medication adherence.
7. Demonstrate that α levels are reasonable.
8. Demonstrate that parameter estimates are not biased.
9. Apply the technique to a dataset from the Adherence in Rheumatoid Arthritis: Nursing Interventions studies (RO1 NR02107 and RO1 NR04554).

2.0 LITERATURE REVIEW

2.1 DEFINITION OF THE J-SHAPED CURVE

Allport (1934) gave several examples of behaviors that followed telic continuums and showed that when the proportion of persons completing various portions of a behavior is plotted, the curve is J-shaped in nature. He described several behaviors that demonstrated this type of distribution including traffic signal obedience, work attendance, and kneeling to pray. In all of these examples, there tends to be a somewhat large proportion of people who do not partake in any portion of the behavior, an even larger proportion that complete the behavior totally and correctly, and a smaller, but substantial, proportion that complete some portion of the behavior, though not completely nor correctly. For example, most drivers when approaching a stop sign will stop completely, while a small number will not stop nor even attempt to slow down. A smaller number will exhibit some behavior in between ranging from applying the brake only enough to light the tail light to slowing significantly but not fully stopping.

The manner in which people take their medication follows a similar pattern. Many people faithfully take their medication completely and correctly. A significant proportion of people fail to take their medication at all. A small, but significant proportion of people intermittently take their medication, either missing doses within a day or missing days at a time. For example, patients on a multiple time per day dosing schedule may fail to take a midday dose.

Indeed in a study on rheumatoid arthritis medication adherence demonstrated that adherence decreased as complexity of dosing schedule increased (Rohay, Dunbar-Jacob, Sereika, & Kwoh, 1996; Dunbar, 1990). Alternatively, patients may be more likely to miss a weekend dose when their routine is disrupted (Aronson, 2006). It follows that adherence measures are measures of conformity and thus, usually result in a J-shaped or U-shaped bimodal distribution. A review by Dunbar-Jacob et al. (1994) demonstrated several examples of adherence data that were J-shaped.

While medication adherence follows the full telic continuum of behavior, the J-shape distribution of these data has impeded full exploration of the whole range of the behavior. Instead data are often collapsed for analysis into an "all or nothing" dichotomous variable. This technique reduces the amount of information used in the available data. Additionally, in studies designed to increase medication adherence, small but significant improvements are often unnoticed because current statistical techniques inherently lack the ability to detect them.

2.2 EARLY MEASURES OF CONFORMITY

An early technique used to characterize the J-shaped distribution and provide a method of comparing two curves was described by Dodd (1949). He introduced the conformity scale that was used to create a single index describing the amount of "conformity" exhibited. However, an inherent flaw with this index, at least as a method of comparing interventions designed to increase adherence, was that several distributions resulted in the same conformity index when one distribution clearly exhibited better adherence. For example, Figure 2.1 displays two distributions that would result in approximately the same conformity index. Though the one on

the left is clearly superior (as more observations are in the upper tail) it has a lower conformity index.

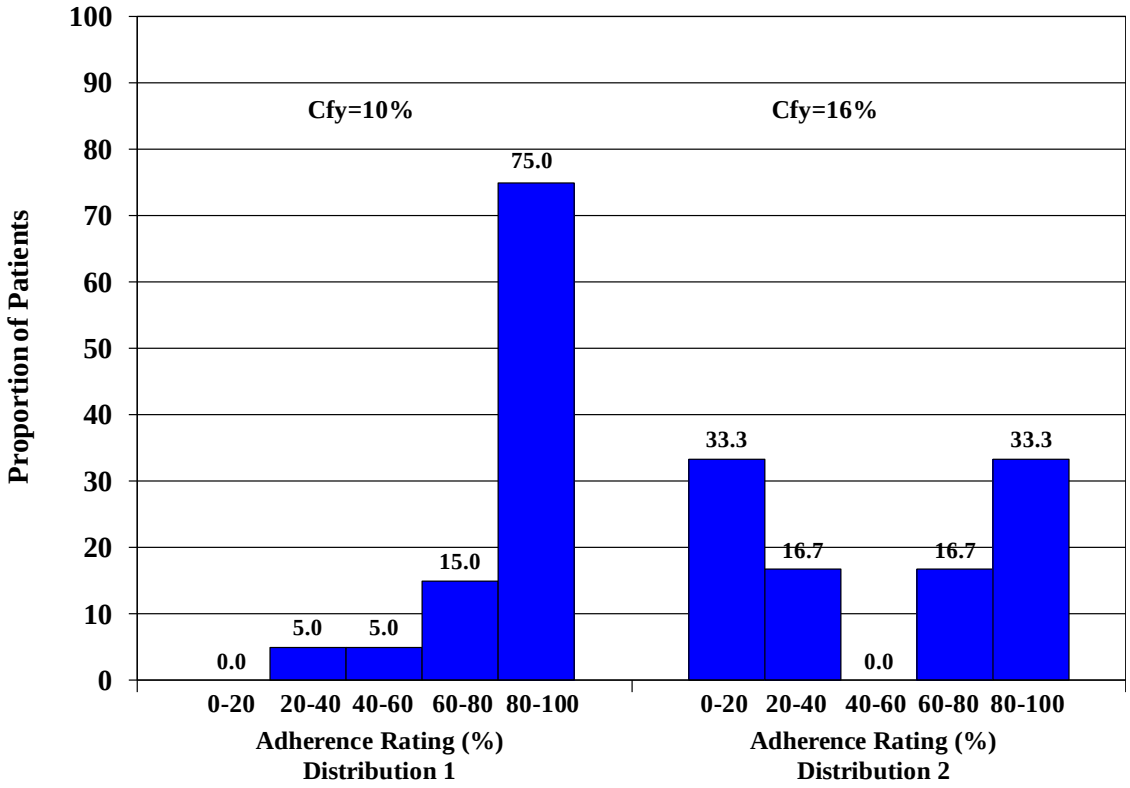


Figure 2.1 Conformity Indices for Two Different J-shaped Distributions

2.3 DEFINITION OF ADHERENCE

In studies on medication adherence, adherence is measured in variety of manners including pill counts, self-report, clinical measures and electronic monitoring. Each measure suffers from bias as a subject can manipulate each in order to increase their apparent adherence level. An assumption is made that the subject took their medication at a constant level over the interval in question.

Pill counts measure adherence by calculating the percentage of pills that are returned at subsequent clinic visits. The problem with this measure is that it is unknown when and if the subject actually took the dose and whether the dosing schedule was followed in an accurate manner, thus maximizing the effectiveness of the dose.

Self-report, via a diary, is another method used to capture the manner in which medication is taken. Again, it relies upon the integrity of the subject to report accurately when they took their medication.

Clinical measures can be used to assess how much medication is in the body at a given point in time. While giving an accurate indication of whether or not the person has taken their medication, these measures can be costly. Additionally, they are specific to the time point of the assay and may not reflect the true nature of the person's medication adherence.

Electronic monitoring makes use of a microchip which allows for the recording of every event (i.e., opening and closing of pill bottle or activation of inhaler). An assumption is still being made, namely that if subject opens the bottle that they are taking a proper dose of medication. However, some forms of manipulations can be identified, such as repeatedly opening and closing the bottle to make it appear that the bottle was opened the correct number of times. There are several advantages to this technology. It increases the amount of information available for calculating adherence. This allows for the computation of adherence over smaller intervals, which eliminates the assumption of a constant adherence rate over a time period. As a result we can better estimate true therapeutic effect of medication by evaluating dose-response relationship over entire interval. Additionally, we can assess changing adherence patterns within intervals and use several summary measures for the interval if necessary.

However, this technology has its disadvantages. First, analyses are potentially more complex as a result of additional data points. For example, adherence measured daily over a two week period may result in 14 measures. Also, missing data must be accounted for as missing data may be due to lack of adherence or the result of monitor malfunction.

Calculating adherence is much more difficult and different methods can lead to varying results. For example, a simple method comparable to a pill count would be to divide the total number of events in interval divided by the number prescribed. A second method would be to calculate the average of daily adherence where daily adherence is number of events on a given day divided by the number prescribed. This method is comparable to the average of several pill counts as it does not account for dosing intervals. Finally, adherence measures can be created for the average of daily adherence that uses dosing intervals. As noted above, the inherent assumption is made that the subject actually consumed the medication at the time the cap was removed and replaced. The observed range of adherence measures is 0% to 100%. It is impossible to be less than 0% adherent and by "folding" or truncating adherence greater than 100% an upper bound of 100% is created. Thus, statistical techniques that may be developed for the analysis of adherence data, must take this into account.

These three methods were compared using screening data for the first RA study (Rohay, Dunbar-Jacob, Sereika, & Kwoh, 1996). The first method resulted in a median adherence rate of 90%. The second method resulted in a median rate of 83%, while the last method resulted in a median adherence rate of 78%. As described above, none of these measures were normally distributed.

While the above methods for calculating adherence results in a single measure, several authors have pointed out that a patient's adherence decreases over time and between clinic visits (Cramer, Scheyer, & Mattson, 1990; Norrell, 1981). Thus, a single summary measure may not adequately describe true adherence.

Furthermore, because adherence is based on a summary measure for a given period, they do not provide any information on how subjects took medication within that time period. Therefore, a measure of variability in adherence within a given interval is lost resulting in a lack of precision. Earlier methods of collecting adherence measures resulted in data that were only relevant for a given period, providing information for a cross section of time. Electronic monitoring allows for continuous measures of adherence. However, data can be aggregated within a specified time period to provide a cross-sectional measurement.

Additionally, the multiple adherence measures can be used "as is" to create a time series of data. This method eliminates the assumption that adherence is constant over a time period by allowing for multiple observations per subject.

Regardless of the manner in which adherence is computed, the resulting distributions tend to be J-shaped. Additionally, the J-shaped nature of adherence measures is not limited to just medication but can be applied to other applications. For example, adherence can be calculated for exercise regimens, screening programs, and the number of treatment sessions attended in a clinical trial. In any case, the measurement still suffers from the inherent problem of being J-shaped in nature.

2.4 CURRENT METHODS OF ANALYSIS

Statistical analysis of adherence data is complex. Electronic monitoring can result in hundreds of data points per subject. Summarizing data may require several measures per interval if adherence is not constant. Several techniques can be used for analysis including using the average of all adherence measures over a given interval. This method is adequate if the measures are relatively constant in interval. If adherence rates are increasing or decreasing over time as several authors have reported, then one measure will not identify this trend in the data. Thus, several measures may be needed resulting in a more complex analysis using a repeated measures technique. Change scores can be calculated for both within and between intervals.

Typically, studies attempting to characterize adherence have reported mean adherence rates. Usually, the standard deviations are also reported. Means and standard deviations may not adequately describe the adherence distribution, especially when J-shaped. While measures of central tendency will provide an understanding of the data, use of the median, mode and range and ideally a histogram may provide the most accurate depiction of the data.

Additionally, reports should include the method used to calculate adherence. As demonstrated earlier calculations using dosing intervals resulted in lower adherence levels. Thus, in order to compare results across studies the methods employed to calculate adherence must be comparable.

Descriptive techniques may be adequate for characterizing adherence rates in a cross-sectional manner however they are inadequate for assessing whether interventions have had an impact on improving adherence rates. It is unlikely that interventions will change very low adherers to adequate adherers, however small increases in adherence levels increase the therapeutic effect of a medication. Analyses using adherence as a continuous variable will be

able to detect every subjects' change. Therefore, analyzing adherence as a continuous variable allows for the detection of small improvements, thus using t-tests, ANOVA, or non-parametric tests such as the Wilcoxon Rank-Sum or the Kolmogorov-Smirnov (K-S) to detect differences in adherence or regression to identify predictors of adherence would seem to be a better technique. Several studies reported average adherence rates and/or analyzed adherence data using established parametric tests. Olivieri et al. (1991) used paired t-tests to compare adherence rates on the first and last day of a 30 day medication regimen and noted they did differ. Jonasson et al. (1999) used t-tests to compare self-reported adherence measures to those based on pill counts of returned medication. Similarly, Matsui et al. (1994) used t-tests to compare adherence rates based on electronic monitoring to those based on pill counts. Finally, McKenney, Munroe and Wright (1992) used t-tests to compare adherence measures comparing two groups of patients – one receiving a medication bottle with a time piece and another receiving standard medication bottles. These techniques require that the underlying distribution of the dependent variable (i.e., adherence) be normally distributed. Though the t-test is robust to unimodal distributions, J-shaped adherence distributions, of course, are not considered to be unimodal or normal, nor are they easily transformed.

One limitation with working with this type of data is that the effect of violating the normality assumption of standard parametric tests has not been explored. One goal of this paper is to explore the effect of this violation and determine its impact on such analyses.

Another technique that has been used to analyze adherence data as a dependent variable is to categorize it into two or more discrete groups. Several studies categorized subjects as either adherent or non-adherent base on a cutoff level. Arbitrary cutoffs of 75% (Gallagher, Viscoli, & Horwitz, 1993; Horwitz, et al., 1990) and 80% (The Coronary Drug Project Research Group,

1980) were common. Bartlett et al. (2002) reported pre- and post-intervention appropriate use adherences rates along with p-values, presumably using χ^2 tests. This rating can be used with practically every type of adherence measure. For example, one study defined three levels of adherence on the basis of self-reported doses of medication used during the study (Pizzo, Robichaud, Edwards, Schumaker, Kramer, & Johnson, 1983) Another study used an arbitrary cutoff for level of serum concentration of the medication (Pachter & Weller, 1993). When treating adherence data in this manner several traditional statistical techniques are appropriate for use. These techniques include χ^2 tests, Fisher's Exact test, and logistic regression. However, when attempting to assess interventions designed to increase adherence, categorizing adherence may not detect changes because improvements may be small and inadequate for subjects to cross the predetermined threshold of adequate adherence.

The use of two or three categories has been noted in the literature. One study (Bramley, Gerbino, Nightengale, & Frech-Tamas, 2006) used several levels of adherence; however that study was not assessing adherence as an outcome. The use of three levels would allow for the comparison of subjects who never (or seldom) take their medication (low, <50%), subjects who always (or almost always) take their medication (high, >80%), and subjects who take their medication intermittently (moderate, 50%-80%). Still, categorizing adherence in this manner results in loss of information as small changes are not accounted for. Thus, only subjects who increase over a predetermined cutoff of "good adherence" will show improvement while those who increase their adherence, but not over the cutoff, will still be labeled non-adherent. Again, the effect of categorizing adherence data is not well understood and will be addressed in this paper. Finally, other transformations, distributional models, and/or statistical methods need to be explored in order to ensure correct statistical techniques are used when analyzing adherence data.

Other distributions, for example a Beta-Binomial distribution, need to be explored in order to characterize the underlying adherence distribution. If a specific distribution can be identified then appropriate statistical techniques can be employed for analysis.

Additionally, other statistical methods such as curve fitting (e.g., Curvefit software) should be explored. Finally, as the J-shaped distribution can be generated from a compound beta distribution, techniques for estimating the parameters can be developed for use in analyzing adherence data. This paper will focus on this technique as a method of analysis.

2.5 THE EXPECTATION-MAXIMIZATION ALGORITHM

We propose that J-shaped distributions, such as those observed for medication adherence, can be approximated by a mixed beta distribution, that is,

$$x_i = \pi Beta(1,\beta) + (1-\pi) Beta(\alpha,1)$$
$$where\ 0 \le \pi \le 1 is the mixing\ paramenter$$

Equation 2.1

However, the parameters of the resulting likelihood function cannot be estimated using well established techniques as the derivates of the likelihood functions do not exist. Pawitan (2001) describes a technique to estimate the parameters of a mixture model when there are constraints on one or more of the parameters. In the case of the equation above, constraints on π make application of the Newton-Raphson technique impossible. By using the Expectation-Maximization (EM) algorithm, the parameters of the model can be estimated. Using this technique, the probability that an observation is from a given component of the mixture

distribution is computed based on initial estimates of the parameters of the mixed model. These probabilities are then standardized and used to create weighted values of the observations. These weighted observations are then used to update the estimate of the parameters. In the case of the beta distribution, the maximum likelihood estimates for the parameters of the beta distribution are used to estimate the α and β parameter (Evans, Hastings, & Peacock, 2000). The process is iterated until convergence.

Furthermore, Pawitan (2001) describes a technique for the estimation of standard errors, allowing for comparisons of parameters in two (or more) different groups. Essentially, the above technique is repeated on either Bootstrap or Jackknife samples and the standard deviation of the bootstraped or jackknifed estimates is used as the standard error. The techniques are similar in that each one samples a given dataset multiple times. The bootstrap samples with replacement a large number of times (at least 200) and produces the estimates. The jackknife creates multiple datasets by dropping the i^{th} observation. The two techniques provide similar results with the jackknife method being more computational efficient (Efron & Tibshirani, 1998).

The EM algorithm has been used for assessing the parameters of other mixed models, and has been used to estimate the parameters of mixed beta distributions (Ji, Wu, Liu, Wang, & Coombes, 2005) for estimating correlation coefficients. However using a mixed beta distribution is a novel approach to estimate the parameters for assessing adherence distributions. As noted above, many researchers use established statistical techniques to analyze adherence data even though the assumptions of these tests were clearly violated. The EMAAPE was used in an attempt to estimate the parameters and their associated standard errors for data from J-shaped distribution.

3.0 METHODS

3.1 DEVELOPMENT OF J-SHAPED DISTRIBUTION

Simulation of the J-shaped distribution was conducted using the composite of two beta functions. A beta distribution with parameters 1 and β produces a decreasing curve, with β controlling the steepness of the resulting curve. A beta distribution with parameters α and 1 produces an increasing curve, with α again controlling the steepness of the curve. (Cassella & Berger, 1990). An advantage of using the beta distribution is that the random deviates are bounded by 0 and 1 and thus is a natural choice for modeling adherence data. A technique, similar to the acceptance-rejection method using exponential envelopes, described by Lange (1998), was used to generate random deviates. Random deviates were generated using Equation 2.1.

Beta(j,k) indicates a random deviate from a beta distribution with parameters j and k. In order to ensure a known proportion of observations were obtained from each component of the mixed distribution, the value of π was fixed.

The value of both β and α in Equation 2.1 above must be greater than 1. A beta distribution with parameters β=1 and α=1 is simply the uniform distribution (Cassella & Berger, 1990) and is not relevant to this problem. Larger values of α and β produce steeper curves which results in distributions with more values at each extreme. Both parameters α and β were

assigned values of 2, 3, 5, 8, 11, and 14. See Figure 3.1 for a comparison of several values of α and β.

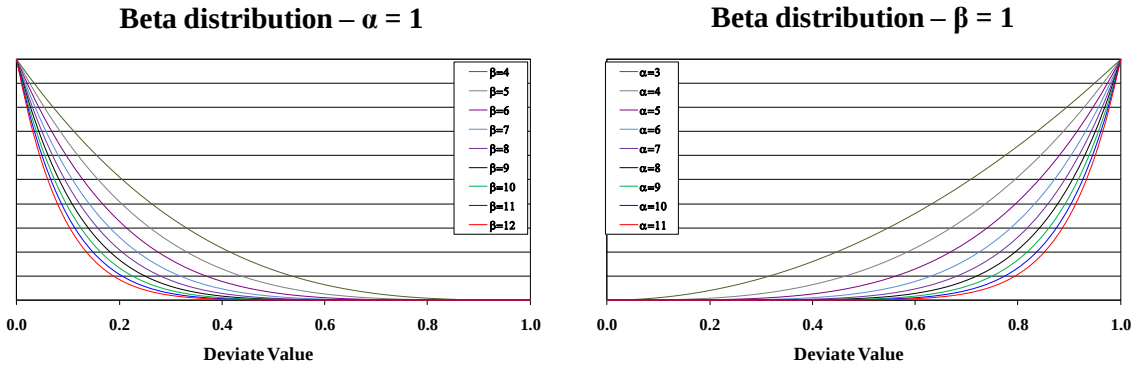


Figure 3.1 Beta Distributions with Parameters (1,β) and (α,1)

Because adherence distributions typically have fewer observations at the lower end, it was hypothesized that the left hand portion should contribute relatively fewer observations than the right hand portion. Thus, the parameter π was fixed at values 0.15 and 0.25.

A random deviate was produced as follows:

1. Generate π times the sample size random deviates from a beta distribution with parameters 1 and β (left hand side of distribution).
2. Generate 1- π times the sample size random deviates from a beta distribution with parameters α and 1 (right hand side of distribution).

Use of the beta distribution resulted in a deviate that was between 0 and 1 and thus had the advantage of not requiring the removal of any out of range observations. A dataset of 100,000 observations was generated for each combination of the parameters for a total of 72

datasets. Histograms were generated and the results visually inspected for appearance of a J-shaped distribution. All programming was completed using SAS 9.1.3 (SAS Institute Inc., 2002-2003).

3.2 TRANSFORMATIONS

In an attempt to meet the assumptions of traditional analytic techniques a variety of transformations were conducted in order to induce normality. A series of Box-Cox transformations (Box & Cox, 1964) were conducted along with several others transformations. Using values of 0.15 and 0.25 for π and the values of 3, 5, 8, 11, and 14[1] for the α and β parameters, 200 datasets of size 200 were created and assessed for each combination of α and β for a total of 10,000 datasets. The transformations that were attempted are shown Table 3.1.

Transformations attempted included Box-Cox transformations, power transformations (square root and squaring which are just special cases of the Box-Cox transformation), inverse, logarithmic and arcsine transformations. For each transformation a Kolmogorov-Smirnov test of normality was conducted. Additionally, histograms of the transformed data were generated for several randomly selected datasets.

[1] α=2 and β=2 were dropped due to the flat nature of distributions including those parameters.

Table 3.1 Transformations Attempted to Induce Normality of J-shaped Data

Class	Function
Box-Cox[1]	$(x^{\lambda}-1)/\lambda$; $\lambda \neq 0$ $\ln(x)$; $\lambda=0$ $-2 \leq \lambda \leq 2$ by increments of 0.1
Square Root Transformation[2]	$(x+0.5)^{1/2}$
	$(x+0.375)^{1/2}$
Square[2]	x^2
Inverse	$(x+1)^{-1}$
Logarithmic	Log_e[2]
	Log_2
	Log_{10}
Arcsine	$\text{Arcsine}\left[\sqrt{x}\right]$ $\text{Arcsine}\left[\sqrt{x}\right]$ for $0 < X < 1$; ¼ for X=0; ¾ for X=1 $\sqrt{\left(n+\frac{1}{2}\right)}\,\text{Arcsine}\left[\sqrt{\left(\frac{x+0.375}{n}+0.75\right)}\right]$

[1] - A total 41 transformations were made
[2] - Special case of Box-Cox transformation

3.3 DEVELOPMENT OF DATASETS USED FOR SIMULATION STUDY

3.3.1 Datasets Used to Evaluate the EMAAPE

Three parameters – π, α, and β, were used to generate datasets for three different sample sizes - 100, 200, and 500. The proportion of observations from the left hand component of the curve, π, assumed the values 0.15 and 0.25. Both β and α assumed five values: 3, 5, 8, 11, and 14. Thus, for each sample size, 50 parameters sets were generated. Two hundred datasets were created for each parameter set. Parameter estimates and their standard error were generated for each using the EMAAPE.

3.3.2 Datasets Used to Assess Type I Error (α Levels)

Using the parameters sets described in section 3.3.1 a second set of datasets were generated and their parameter estimates and associated standard errors were generated using the EMAAPE. These datasets were paired with those created above such that each pair was of the same size and created using the same parameters and an arbitrary dataset number. One dataset of the pair was designated as "Group I" and the second was designated as "Group II".

3.3.3 Datasets Used to Assess Type II Error (β Levels)

The datatsets generated in sections 3.3.1 were again used to assess the Type II error rate. Datasets for each parameter set were compared with datasets for every other parameter set. Datasets were again paired based on an arbitrary dataset number with one dataset of the pair designated as "Group I" and the second designated as "Group II". For each sample size a total of 245,000[2] comparisons were made. In order to provide a sense of "effect size" the sum of difference between the distributions was computed using the following steps:

1. Group the data for each dataset into deciles.
2. Determine the "better" group based on the lower proportion of observations in the first decile (if equal move to the next decile until one distribution is deemed "better").
3. Determine at which decile the distributions cross.

[2] For the first parameter set, 49 comparisons were made. For the second, 48 comparisons, the third 47, etc. Thus, there were 49+48+47+…+1 comparisons = 1225 comparisons for each of the 200 datasets for a total of 245,000.

4. Until the distributions cross, compute the absolute difference between each decile.
5. After the distributions cross, compute the difference between each decile.
6. The Sum of the Difference in Distributions (DD_SUM) is the sum of the differences computed in steps 4 and 5.

3.4 THE E-M ALGORITHM

Pawitan (2001) describes a technique to estimate the parameters of a mixture model when there are constraints on one or more of the parameters. The assumed model for data from a J-shaped distribution is a beta mixture model as shown in Equation 2.1.

In the case of the Equation 2.1, constraints on π make application of the Newton-Raphson technique impossible to perform. By using the Expectation-Maximization (EM) algorithm the parameters of the model can be estimated. Furthermore, he describes a technique for the estimation of standard errors, allowing for comparisons of parameters in two (or more) different groups.

While use of the E-M algorithm has been used for assessing the parameters of other mixed models, the EMAAPE is novel in its use for estimating the parameters of a mixed beta distribution characterizing adherence distributions. As noted above, many researchers use statistical techniques to analyze adherence data even though the assumptions of these tests are clearly violated.

3.4.1 Estimation of Parameters for Composite Beta Distribution

The EMAAPE was used to produce parameter estimates for data coming from the beta mixture model (Equation 2.1). The parameters estimated were π, β, and α. The remaining parameters of the beta distributions were assumed to be one for the purpose of generating the strictly increasing or decreasing curves and thus, the algorithm assumed these parameters were fixed at one. The steps necessary to produce the estimates are:

1. Estimation step

   a. Compute

$$\hat{r}_i = \frac{\pi_1^{i-1} p(x_i \mid 1, \beta_1^{i-1})}{\sum_k \pi_k^{i-1} p(x_i \mid \alpha_k^{i-1}, \beta_k^{i-1})}, k = 1,2; \alpha_1^{i-1} = 1$$

Equation 3.1

   the conditional probability that the observation belongs to the i^{th} component of the mixed distribution.

2. Maximization step

   a. Update $\hat{\pi}, \hat{\beta}_1, \hat{\alpha}_2$,

   1. $\pi_1^1 = \frac{\sum_i \hat{r}_{i1}}{n}, where \hat{r}_{i1} = \frac{p(x_i; \hat{b}_1, 1)}{p(x_i, \hat{b}_1 1) + p(x_i; 1, \hat{a}_2)}$

Equation 3.2

   2. $\beta_1^1 = \frac{\sum \rho_{i1}}{\sum \rho_{i1} \ln(1 - x_i)}, p_{i1} = \frac{p(x_i; \hat{b}_1, 1)}{p(x_i, \hat{b}_1 1) + p(x_i; 1, \hat{a}_2)}$

Equation 3.3

3. $$\alpha_2^1 = \frac{\sum \rho_{i2}}{\sum \rho_{i2} \ln(x_i)}, \; p_{i2} = \frac{p(x_i;1,\hat{a}_2)}{p(x_i,\hat{b}_1 1) + p(x_i;1,\hat{a}_2)}$$

Equation 3.4

b. Compute the log-likelihood (Ji, Wu, Liu, Wang, & Coombes, 2005):

$$l(\beta_1,\alpha_2,\pi) = \sum_{i=1}^{n} \pi[\ln(\pi) + \ln(f_1(x_i \mid 1,\beta_1)] + (1-\pi)[\ln(1-\pi) + \ln(f_2(x_i \mid \alpha_2,1)]$$

Equation 3.5

c. Iterate until convergence using a criterion of a difference less than 0.0001 for log-likelihood or until a maximum of 300 iterations.

The maximum likelihood estimates for the beta distribution are given by Evans (2000, p. 41) and are found by solving the following simultaneous equations:

$$\psi(\hat{\alpha}) - \psi(\hat{\alpha} + \hat{\beta}) = n^{-1} \sum_{i=1}^{n} \ln(x_i)$$

$$\psi(\hat{\beta}) - \psi(\hat{\alpha} + \hat{\beta}) = n^{-1} \sum_{i=1}^{n} \ln(1 - x_i),$$

$$where \; \psi(\bullet) \; is \; the \; digamma \; function$$

Equation 3.6

The digamma function satisfies the recursion relation (Gautschi, 1997):

$$\psi(x+1) = \psi(x) + \frac{1}{x}$$

Equation 3.7

Thus, the maximum likelihood estimates are given by the following formulas:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(p_i)}{\sum_{i=1}^{n}(p_i)\ln(1-x_i)}$$

$$\hat{\alpha}_2 = \frac{\sum_{i=1}^{n}(1-p_i)}{\sum_{i=1}^{n}(1-p_i)\ln(x_i)}$$

Equation 3.8

The EMAAPE was programmed using SAS version 9.1.3 (SAS Institute Inc., 2002-2003). See Appendix A for the macro used to estimate the parameters of the beta mixture model.

3.4.2 Estimation of Standard Errors of Parameters for Composite Beta Distribution

Efron (1982) describes a blocked jackknifing technique for deriving the standard error estimates of the parameters produced using the E-M algorithm. The following steps were used to derive the standard error estimates for each parameter:

1. Using original dataset of size n, estimate the parameters of the beta mixture model.
2. Create datasets of size n-h by removing the first h observations where h is equal to the sample size divided by 100 (e.g., for n=100, eliminate 100/100=1 observation per replicate; for n=200, eliminate 200/100=2 observations per replicate; etc.).
3. Estimate the parameters of the beta mixture model using the jackknife sample.

4. Repeat steps 2 and 3 an additional 99 times with the i^{th} block of observations removed.
5. For each parameter, estimate the standard error as follows:

$$\hat{\sigma}_{\hat{\theta}_j} = \sqrt{\frac{n-1}{n}\sum(\hat{\theta}_{(i)} - \hat{\theta}_j)^2}$$

Equation 3.9

See Appendix B for the SAS macro used to estimate the standard errors. These macros were run on each of the datasets described in sections 3.3.1 and 3.3.2 and the results were used to assess the performance of the algorithm, the type I error rate, and the type II error rate.

3.4.3 Validation of Parameter Estimates

Efron (1982) also states that jackknifing can be used to assess bias of the parameter estimates by subtracting the mean of the parameter estimates from the jackknife samples from the known parameter. However, Mooney (1993) points out that estimating bias estimate from only one sample may be inadequate as the amount of random variability that exists along with bias in the bootstrap bias estimator may be impossible to determine. Thus, bias was assessed for each sample size by comparing the means of the parameter estimates from the 10,000 datasets per sample size described in section 3.3.1 to the parameters used to generate the datasets.

Performance of the EMAAPE was assessed by counting the number of datasets in which the likelihood did not converge. Additionally, the number of datasets in which the maximum likelihood was attained on the first iteration was also tallied. Results assessing parameter estimates, Type I, and Type II error rates excluded these datasets.

3.5 ASSESSMENT OF TYPE I ERROR

The EMAAPE was run on each dataset described in sections 3.3.1 and 3.3.2. The parameters and their associated standard errors were calculated as outlined in sections 3.4.1 and 3.4.2. Datasets of equal size, generated using the same set of parameters, were paired using an arbitrary dataset number. For each pair of datasets, the parameters and their standard errors were compared to determine the proportion of times the two datasets were deemed different based on an overall alpha level of 0.05. For each dataset pair, a t-score based on the difference in the parameter estimates and their associated standard errors was calculated as follows:

$$t = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{se_s}},$$

where θ is a given parameter,

$$se_S = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \text{ and}$$

$$s_i^2 = n \cdot se_i^2$$

Equation 3.10

In Equation 3.10, se_s is the Satterthwaite approximation (Rosner, 1990) of the pooled standard error which will take into account unequal variances as well as unequal sample sizes. In cases where the variances are equal, this method provides pooled variance estimates which are very similar to those of the standard method of calculating the pooled variance estimates. Thus, during simulations no attempts were made to assess differences in the standard errors and the Satterthwaite approximation was used throughout.

The specific hypothesis tested was:

H_0: $\pi_1 = \pi_2$ and $\beta_1 = \beta_2$ and $\alpha_1 = \alpha_2$

H_A: At least one of the parameters was not equal.

A statistically significant difference was noted when at least one of the parameters was deemed significant based on the observed p-value. This criterion was used, as opposed to basing results on just one parameter, as it was possible that a portion or portions of each curve might have had very similar characteristics (i.e., derived from the same beta distribution) but differ on only one parameter in the model (e.g., the proportion of data points generated from a specific beta distribution). However, in order to maintain an overall alpha level of 0.05, alpha was partitioned as follows: 1) for comparing α_1 v. α_2 – p $\leq$ 0.035; 2) for comparing β_1 v. β_2 – p $\leq$ 0.005; and 3) for comparing π_1 v. π_2 – p $\leq$ 0.01. The majority of the significance level was given to the α parameter as it was hypothesized that this would be estimated most accurately given the larger proportion of data points from the right hand portion of the curve. Additionally, the proportion of observations from the generated data at or above 0.8 (an indicator of good adherence) increased with increasing values of this parameter. This was not seen for the β parameter (see Figure 3.2). While the proportion of observations at or above 0.8 decreased with increasing values of the mixing parameter π, it was only given slightly more weight than the β parameter as it is likely in actual trials that when it is very similar, differences may exist between two curves that might be missed if too much of the significance level is allocated to this parameter. The type I error rate, α, was the proportion of times a pair of datasets were deemed statistically different on a least one parameter allocating alpha as described above.

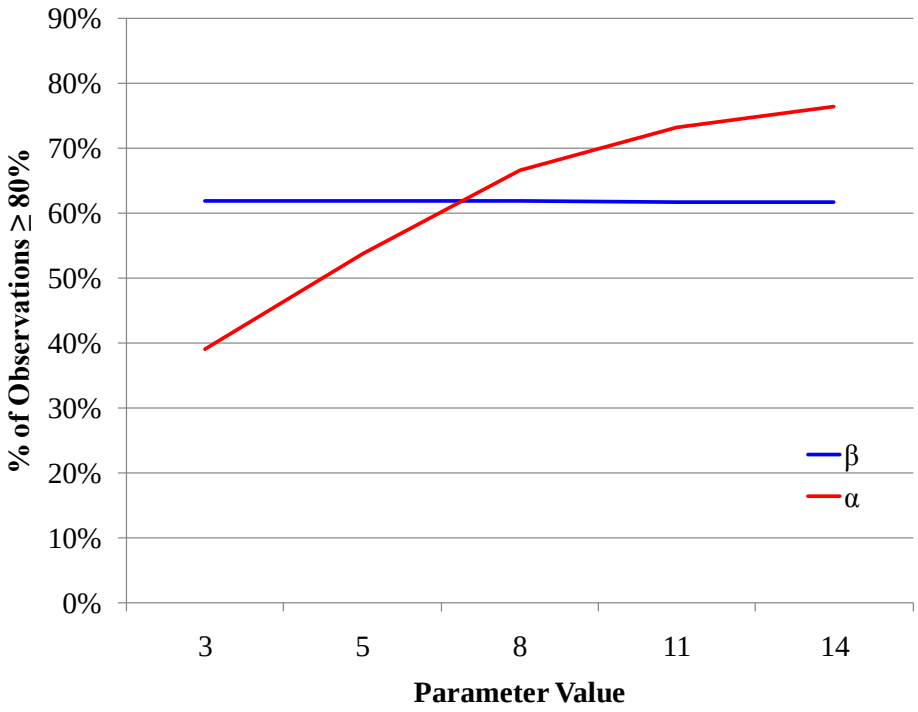


Figure 3.2 Proportion of Generated Observations ≥ 0.8 by Parameter

Additionally, we examined the effect that relaxing the criteria for establishing statistical significance had on the Type I error rate. By relaxing the criteria we hoped to improve the power of the EMAAPE. Thus, we evaluated the datasets using the following criteria: 1) for comparing α_1 v. $\alpha_2 - p \leq 0.04$; 2) for comparing π_1 v. $\pi_2 - p \leq 0.04$; and 3) for comparing β_1 v. $\beta_2 - p \leq 0.02$.

3.6 ASSESSMENT OF TYPE II ERROR

3.6.1 Comparison of Datasets Using the EMAAPE

An empirical power assessment of the EMAAPE was conducted by determining the proportion of datasets that were deemed statistically different among all the comparisons in which at least one parameter was known to differ. For each sample size, each of the 245,000 pairs of datasets discussed in section 3.3.3 were compared. In order to assess whether parameter estimates from two datasets created using different parameters were deemed statistically different, a t-score for the difference in the parameters was created using Equation 3.10.

P-values were then generated based on the t-statistic. Because it is possible that the same parameter may be used to generate portions of both curves, the distributions were deemed statistically significantly different if the p-value for any one of the given parameters was significant allocating alpha as above, separately using both sets of criteria. Again, the hypothesis tested was as follows:

H_0: $\pi_1 = \pi_2$ and $\beta_1 = \beta_2$ and $\alpha_1 = \alpha_2$

H_A: At least one of the parameters was not equal.

As above, the parameters and their standard errors were computed and compared to determine the proportion of times they were deemed different. The type II error rate, β, was the proportion of times that the algorithm failed to detect a statistical difference on a least one parameter.

Results that were statistically significant had different interpretations depending on the specific parameter that was deemed different. A significant difference in the parameter π would indicate that one group, that with the larger value of this parameter, was inferior as more of the observations occurred in the lower ranges of the adherence distribution. An example of two such distributions is shown in Figure 3.3. Similarly, a significant difference in the parameter β would indicate that one group, again the one with the larger value of this parameter, was inferior as more of the observations would be clustered at very low levels of the adherence distribution. An example of two such distributions is shown in Figure 3.4. And finally, a significant difference in the parameter α would indicate that one group, that with the larger value of this parameter, was superior as more of the observations would occur at very high levels of the adherence distribution. An example of two such distributions is shown in Figure 3.5.

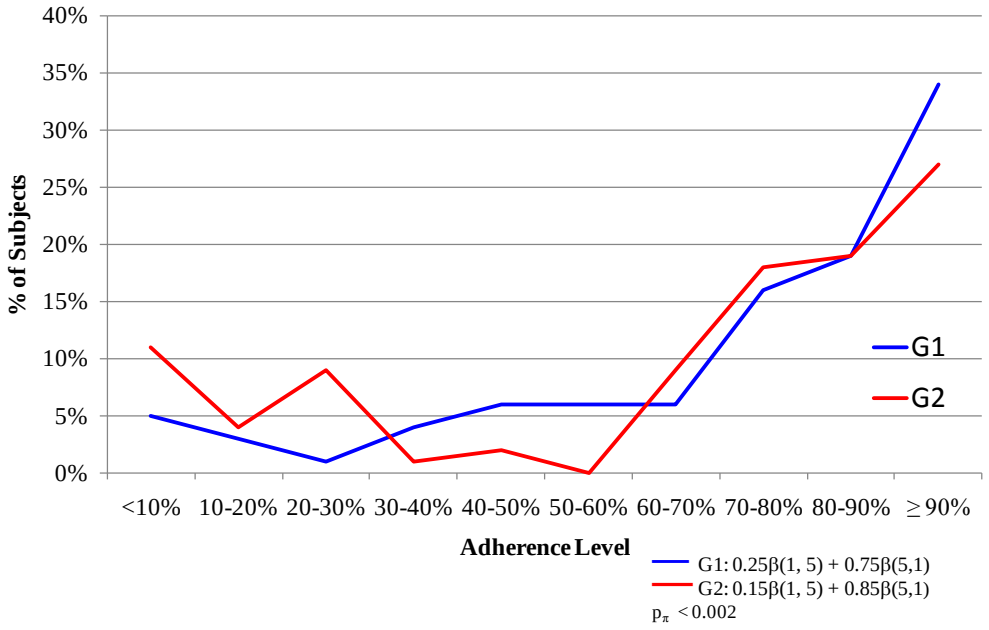


Figure 3.3 Statistically Different Distributions Based on Mixing Parameter π

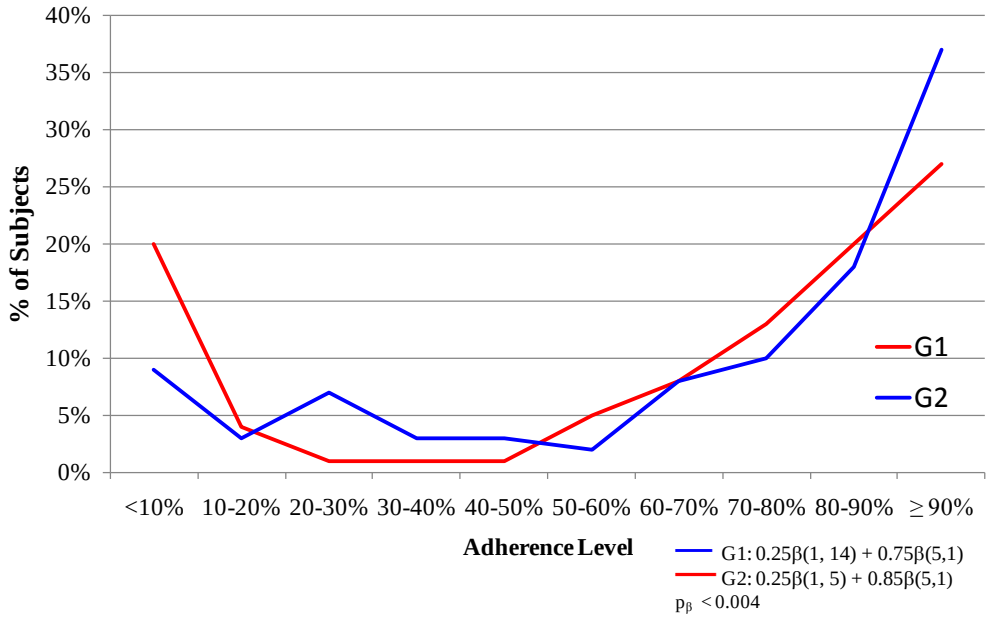


Figure 3.4 Statistically Different Distributions Based on Parameter β

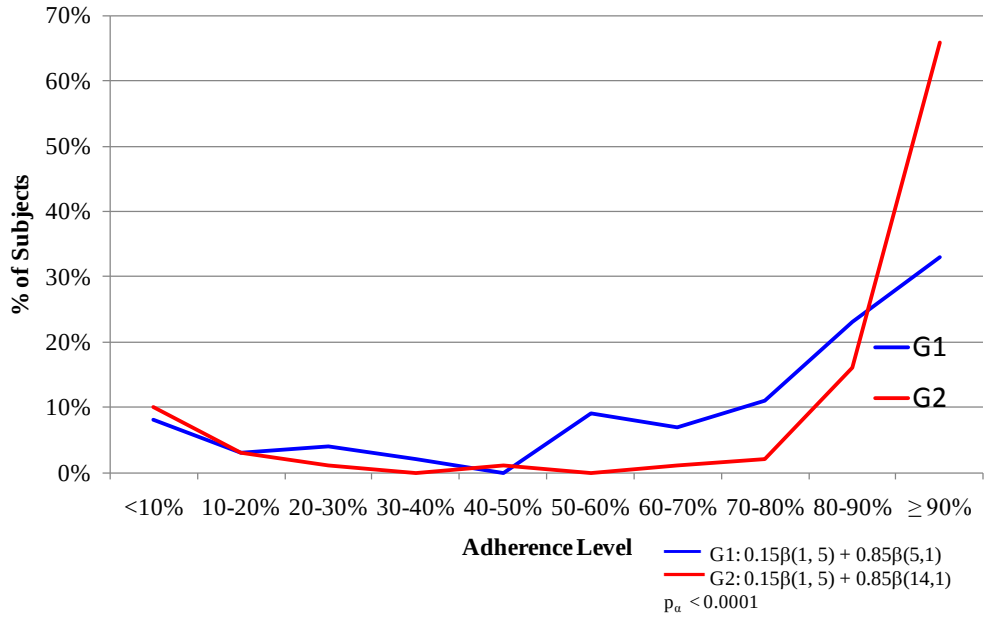


Figure 3.5 Statistically Different Distributions Based on Parameter α

3.6.2 Comparison of Datasets Using Established Techniques

In order to compare results based on the EM Algorithm, the 245,000 pairs of datasets per sample size discussed in section 3.3.3 were compared using the following established techniques: 1) t-test; 2) χ^2 test of association for dichotomized variables (dichotomized as < 0.8 v. ≥ 0.8); 3) Wilcoxon Rank-Sum test; 4) Kolmogorov-Smirnov test; 5) five category (quintiles) logistic regression; and 6) ten category (deciles) logistic regression. Datasets that did not converge or had divergent likelihoods, and those which did not have standard error estimates due to convergence or likelihood issues were excluded from comparisons. For each method the proportion of times datasets were deemed statistically significantly different was noted and compared to results from the EMAAPE.

3.7 ANALYSIS OF RHEUMATOID ARTHRITIS STUDY DATA

The proposed technique was used to evaluate datasets from two "Adherence in Rheumatoid Arthritis: Nursing Interventions" studies (RO1 NR02107 and R01 NR04554**)**. These clinical trials were designed to increase adherence to a medication regime used to treat rheumatoid arthritis. The studies are described in Dunbar-Jacob et al. (1995) but briefly the two trials were similar in methodology and each trial had two interventions designed to improve medication adherence in common: 1) a multi-component behavioral intervention; and 2) a usual care component. Upon completion of the six month follow-up, subjects in the multiple component behavioral intervention were split into two groups: a) a maintenance component; and b) a non-maintenance component.

Two analyses were conducted using these data. The first was conducted on baseline data from the second study (R01 NR04554) and was designed to assess the performance of the EMAAPE for estimating parameters of data collected in an actual study assessing adherence rates. Baseline data from only this study was used as it included a group of "good adherers", thus ensuring a full range of the distribution. The second analysis was conducted to assess the performance of the EMAAPE as a tool for distinguishing two different adherence distributions. Here, data were combined from the two studies in order to create groups of adequate size. Analyses consisted of adherence measured immediately prior to the six month follow-up and thus all subjects receiving the behavioral intervention were included in the treatment group. An electronic monitor was used to capture the opening and closing of the pill bottle signaling the presumptive ingestion of the medication. Adherence was calculated by first computing the daily adherence rate over a 28 day period and the final measure was calculated as the average of the daily measures over the 28 day period. Only subjects taking medication on a daily basis were included in analyses. Subjects with daily adherence measures beyond 100% had their measure set to 100%[3] (as they received 100% of the therapeutic effect) and subjects who dropped out prior to the six month evaluation had their adherence measure imputed, using the adherence measure for the most recent 28 day period prior to the six month intervention. The EMAAPE was used to estimate the parameters and their standard errors for the adherence distribution for each intervention. The formula for the t-statistic in section 3.5 (Equation 3.10) was used to make pairwise comparisons to determine if the two groups differed. These results were contrasted with those obtained using established statistical techniques.

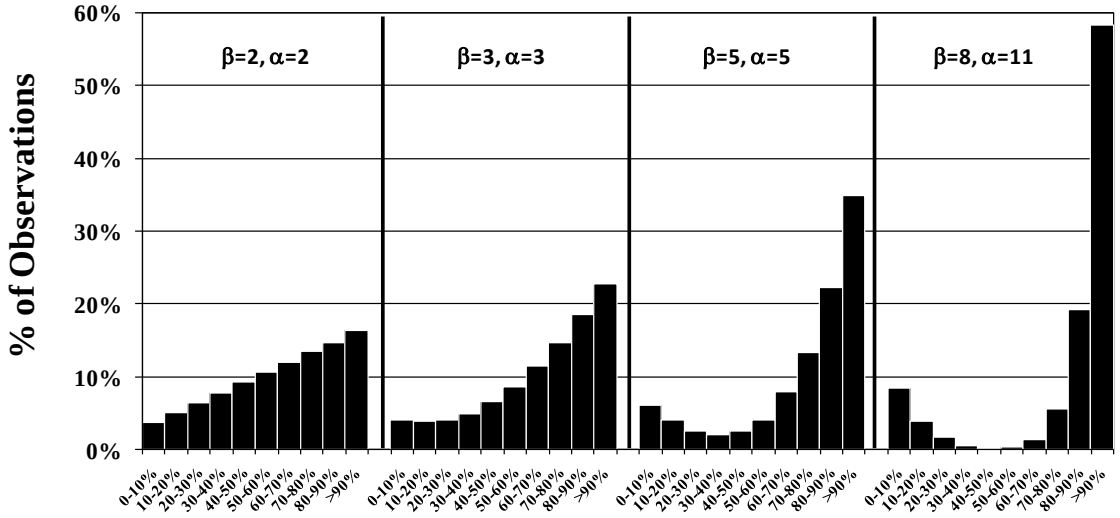
[3] Values of 0 and 100 were replaced by 0.001 and 0.999 for computational reasons

4.0 RESULTS

4.1 DEVELOPMENT OF J-SHAPED DISTRIBUTION

In order to demonstrate that J-shaped distributions could be well-approximated by composite beta distributions, seventy-two datasets of size 100,000 were created as described in section 3.1 and histograms were generated. The large number of observations were generated to ensure the accuracy of the distribution. Figure 4.1 displays the histograms for selected datasets. Histograms of all datasets are shown in Appendix C. In general, datasets where either β or α was two resulted in a poor approximation to a J-shaped curve. As a result, datasets using values of two for β and/or α were not explored further. When the value of both β and α was larger ($\geq$ 11) the distributions demonstrated fewer observations toward the middle. Despite this issue, and because when paired with a smaller value of the other parameter, larger values of one parameter demonstrated adequate results, these values were included in further analyses. Values of β and α that resulted in the best approximation to a J-shape were the values of β=5 and α=5 or α=3. Finally, the value of π did not result in drastic differences in the shapes of the distributions other than the obvious difference in the proportion observed from each component.

π=0.15



π=0.25

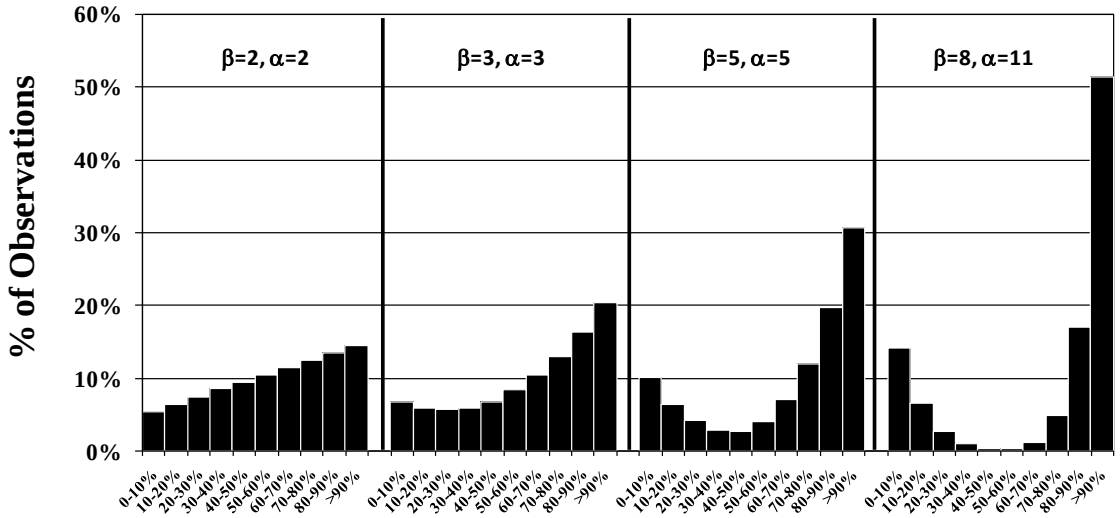


Figure 4.1 Histogram of 100,000 Random Deviates from Selected Mixed Beta Distributions

4.2 TRANSFORMATIONS

In an attempt to meet the assumptions of traditional analytic techniques a variety of transformations were conducted in order to induce normality. A series of Box-Cox transformations were conducted along with several others transformations. Using values of 0.15 and 0.25 for π and the values of 3, 5, 8, 11, and 14[4] for the α and β parameters, 200 datasets of size 200 were created and assessed for each combination of π, α, and β for a total of 10,000 datasets. The transformations that were attempted are shown Table 3.1. None of the transformations attempted resulted in normally distributed data. The distributions of the transformed data for the parameter set π=0.15, β=5 and α=5 are shown in Figure 4.2. Clearly none of the transformations approached normality.

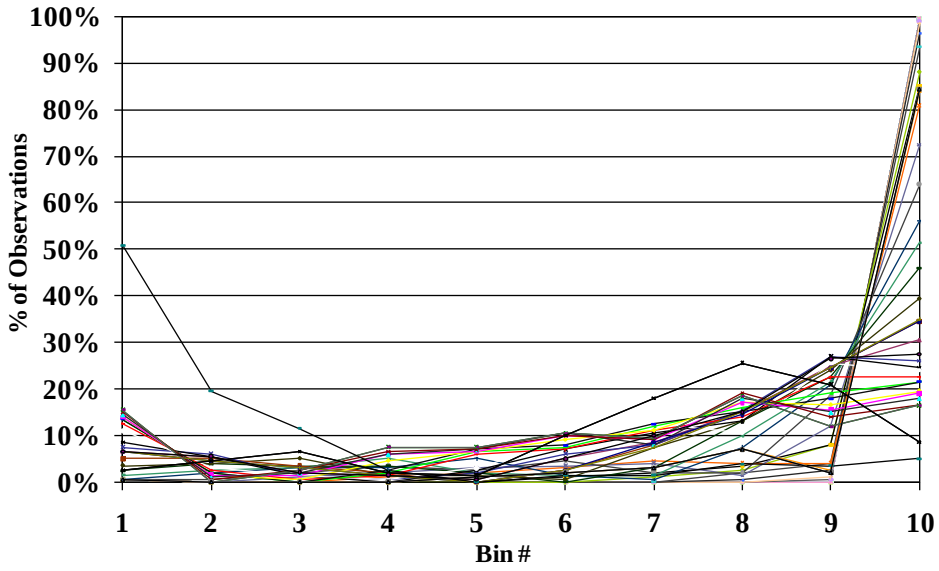


Figure 4.2 Distributions of Transformed Variable: Parameter Set π=0.15, β=5, α=5

[4] α=2 and β=2 were dropped due to the flat nature of distributions including those parameters.

4.3 VALIDATION OF PARAMETER ESTIMATES

Bias was assessed for each sample size by comparing the means of the parameter estimates from the 10,000 datasets per sample size described in section 3.3.1 to the parameters used to generate the datasets. Additionally, the number of datasets that did not converge, and/or attained their maximum likelihood value using initial parameter estimates were counted. Across sample sizes, 600 datasets were generated for each set of parameters. Of the 30,000 datasets generated 1,198 (4.0%) either did not converge or attained a maximum log-likelihood value using the initial parameter estimates. Of these datasets, 527 (44.0%) failed on both. Table 4.1 displays the convergence statistics across all sample sizes for each parameter combination.

Table 4.1 Number (%) of Datasets Not Converging for Each Parameter Combination: All Sample Sizes

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	331 (55.2%)	163 (27.2%)	20 (3.3%)	3 (0.5%)	2 (0.3%)
	5	143 (23.8%)	11 (1.8%)	1 (0.2%)	-	-
	8	17 (2.8%)	1 (0.2%)	-	-	-
	11	4 (0.7%)	-	-	-	-
	14	1 (0.2%)	-	-	-	-
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	349 (58.2%)	73 (12.2%)	2 (0.3%)	-	-
	5	71 (11.8%)	4 (0.7%)	-	-	-
	8	2 (0.3%)	-	-	-	-
	11	-	-	-	-	-
	14	-	-	-	-	-

Note: 600 datasets were generated per parameter combination

By far, the parameter combination that performed the poorest was β=3 and α=3. For both π=0.15 and π=0.25, over half the datasets (55.2% and 58.2% respectively) failed to converge (see Table 4.1). Additionally, when π was 0.15 and either β was 3 or α was 3, and the other parameter was 5, about a quarter of the datasets failed to converge (27.2% for β=3 and α=5 and 23.8% for β=5 and α=3). However, results were improved when π was 0.25 for these parameter combinations. Only 12.2% of the datasets failed to converge when β was 3 and α was 5, while 11.8% failed to converge when β was 5 and α was 3. All other parameter combinations performed well with no more that 3.3% of the datasets failing to converge. Results were consistent across sample sizes with approximately 4% of the datasets failing to meet convergence and/or likelihood criteria with similar patterns observed for parameter combinations.

For a sample size of 100, 10,000 datasets were generated – 200 per parameter combination. A total of 408 datasets (4.1%) failed to converge or attained its maximum likelihood using initial estimates. Of these datasets, 167 (40.9%) failed on both. The results are shown in Table 4.2.

The parameter combination that performed the poorest was β=3 and α=3. For both π=0.15 and π=0.25, almost half of the datasets (48.5% and 41.5% respectively) failed to converge. Additionally, when π was 0.15, β was 3, and α was 5, almost a third (31.0%) of the datasets failed to converge. When β was 5 and α was 3, almost one in five (19.0%) of the datasets failed to converge. When π was 0.25, 16.0% and 19.5% of the datasets failed to converge for these combinations of β and α, respectively. All other parameter combinations performed reasonably well with no more that 9.0% of the datasets failing to converge.

Table 4.2 **Number (%) of Datasets Not Converging for Each Parameter Combination: Sample Size = 100**

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	97 (48.5%)	62 (31.0%)	18 (9.0%)	3 (1.5%)	-
	5	38 (19.0%)	10 (5.0%)	1 (0.5%)	-	-
	8	13 (6.5%)	1 (0.5%)	-	-	-
	11	3 (1.5%)	-	-	-	-
	14	1 (0.5%)	-	-	-	-
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	83 (41.5%)	32 (16.0%)	2 (1.0%)	-	-
	5	39 (19.5%)	3 (1.5%)	-	-	-
	8	2 (1.0%)	-	-	-	-
	11	-	-	-	-	-
	14	-	-	-	-	-

Note: 200 datasets were generated per parameter combination

For a sample size of 200, 390 datasets (3.9%) failed to converge or attained it's maximum likelihood using initial estimates. Of these datasets, 202 (51.8%) failed on both. The results are shown in Table 4.3.

Again, the parameter combination that performed the poorest was β=3 and α=3. For both π=0.15 and π=0.25, over half of the datasets (58.0% and 58.5% respectively) failed to converge. Additionally, when π was 0.15 and either β was 3 or α was 3, and the other parameter was 5, about a quarter of the datasets failed to converge (24.5% for β=3 and α=5 and 27.0% for β=5 and α=3). However, results were improved when π was 0.25 for these parameter combinations. Only 12.0% of the datasets failed to converge when β was 3 and α was 5, while

9.5% failed to converge when β was 5 and α was 3. All other parameter combinations performed well with no more that 2.0% of the datasets failing to converge.

Table 4.3 Number (%) of Datasets Not Converging for Each Parameter Combination: Sample Size = 200

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	116 (58.0%)	49 (24.5%)	2 (1.0%)	-	2 (1.0%)
	5	54 (27.0%)	1 (0.5%)	-	-	-
	8	4 (2.0%)	-	-	-	-
	11	1 (0.5%)	-	-	-	-
	14	-	-	-	-	-
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	117 (58.5%)	24 (12.0%)	-	-	-
	5	19 (9.5%)	1 (0.5%)	-	-	-
	8	-	-	-	-	-
	11	-	-	-	-	-
	14	-	-	-	-	-

Note: 200 datasets were generated per parameter combination

For a sample size of 500, 400 datasets (4.0%) failed to converge or attained it's maximum likelihood using initial estimates. Of these datasets, 158 (39.5%) failed on both. The results are shown in Table 4.4.

Again, the parameter combination that performed the poorest was β=3 and α=3. For π=0.15, over half of the datasets (59.0%) failed to converge while almost three-quarters (74.5%) failed to converge when π was 0.25. Additionally, when π was 0.15 and either β was 3 or α was 3, and the other parameter was 5, about a quarter of the datasets failed to converge (26.0% for

β=3 and α=5 and 25.5% for β=5 and α=3). However, results were improved when π was 0.25 for these parameter combinations. Only 8.5% of the datasets failed to converge when β was 3 and α was 5, while 6.5% failed to converge when β was 5 and α was 3). Every dataset converged for the remaining parameter combinations at this sample size.

Table 4.4 Number (%) of Datasets Not Converging for Each Parameter Combination: Sample Size = 500

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	118 (59.0%)	52 (26.0%)	-	-	-
	5	51 (25.5%)	-	-	-	-
	8	-	-	-	-	-
	11	-	-	-	-	-
	14	-	-	-	-	-
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	149 (74.5%)	17 (8.5%)	-	-	-
	5	13 (6.5%)	-	-	-	-
	8	-	-	-	-	-
	11	-	-	-	-	-
	14	-	-	-	-	-

Note: 200 datasets were generated per parameter combination

Table 4.5 displays the parameter estimates for π by sample size. The EMAAPE estimated this parameter very well with the means and medians equaling the true value of the parameter regardless of sample size. As expected the range of values, regardless of the true value, decreased with increasing sample sizes.

Table 4.5 Estimates of Parameter π by Sample Size

True Value	Sample Size	# of Datasets	Mean	Median	SD	Min	Max
0.15	100	4,753	0.150	0.150	0.024	0.000	0.429
	200	4,771	0.150	0.150	0.017	0.029	0.303
	500	4,779	0.149	0.150	0.012	0.045	0.229
0.25	100	4,839	0.249	0.250	0.030	0.015	0.496
	200	4,839	0.249	0.250	0.022	0.054	0.416
	500	4,821	0.249	0.250	0.014	0.109	0.367

Table 4.6 displays the parameter estimates for β by sample size. In general, the average estimated values of this parameter were somewhat larger than the true value. Median values were somewhat improved though still larger. Ranges were often quite large. These issues were more often noted for sample sizes of 100 and for smaller values of β. For this reason, 80% trimmed means were generated which showed improvement, though they were still somewhat inflated.

Table 4.6 Estimates of Parameter β by Sample Size

True Value	Sample Size	# of Datasets	Mean	Median	SD	Min	Max	Trimmed Mean (80%)
3	100	1,703	4.44	3.42	5.21	0.51	103.88	3.51
	200	1,690	3.76	3.18	2.78	0.87	56.21	3.24
	500	1,664	3.35	3.12	1.23	1.55	24.44	3.13
5	100	1,909	5.92	5.26	3.65	1.39	83.03	5.32
	200	1,925	5.39	5.11	1.83	1.55	30.58	5.16
	500	1,936	5.11	5.03	0.93	2.55	14.23	5.05
8	100	1,984	8.64	8.14	3.45	2.60	85.16	8.22
	200	1,996	8.33	8.09	2.63	3.14	91.65	8.13
	500	2,000	8.14	8.05	1.18	3.94	15.65	8.07
11	100	1,997	11.68	11.14	3.68	1.86	43.98	11.24
	200	1,999	11.40	11.16	2.37	3.93	25.00	11.20
	500	2,000	11.13	11.06	1.34	6.53	16.74	11.08
14	100	1,999	14.80	14.20	4.17	4.16	39.77	14.31
	200	2,000	14.40	14.14	2.78	6.52	33.57	14.22
	500	2,000	14.14	14.02	1.70	8.45	23.42	14.06

Table 4.7 displays the parameter estimates for α by sample size. The EMAAPE estimated this parameter very well with the means and medians very near the true value of the parameter regardless of sample size. The range of values, though small relative to those observed for the parameter β, decreased with increasing sample sizes, regardless of the true value.

Table 4.7 Estimates of Parameter α by Sample Size

True Value	Sample Size	# of Datasets	Mean	Median	SD	Min	Max
3	100	1,724	3.04	3.01	0.58	1.51	6.94
	200	1,689	2.99	2.98	0.40	1.70	5.31
	500	1,669	2.99	3.00	0.26	2.06	4.16
5	100	1,892	5.11	5.05	0.79	2.80	9.85
	200	1,925	5.06	5.03	0.54	3.36	8.65
	500	1,931	5.02	5.01	0.33	3.73	6.70
8	100	1,979	8.12	8.04	1.07	4.79	13.70
	200	1,998	8.09	8.04	0.73	5.37	11.22
	500	2,000	8.04	8.01	0.47	6.41	10.25
11	100	1,997	11.20	11.11	1.37	7.27	19.77
	200	2,000	11.09	11.02	0.97	8.34	15.54
	500	2,000	11.04	11.03	0.61	8.98	13.53
14	100	2,000	14.28	14.16	1.72	9.68	24.74
	200	1,998	14.07	14.03	1.17	10.83	20.69
	500	2,000	14.05	14.03	0.73	11.64	17.72

4.4 ASSESSMENT OF TYPE I ERROR

Assessment of the Type I error rate was conducted using the parameter estimates and standard errors generated using the EMAAPE for each of datasets described in sections 3.3.1 and 3.3.2. The parameters and their associated standard errors were calculated as outlined in sections 3.4.1 and 3.4.2. For each sample size, 10,000 datasets were created that contained data generated from the same parameter set. Half of the observations were designated as "Group I" and half

designated as "Group II" and t-statistics were generated as described in section 3.5. In order to maintain an overall alpha level of 0.05, datasets were deemed statistically different using the following strict criteria: if the p-value comparing the α parameters was less than or equal to 0.035, or if the p-value for the π parameter was less than or equal to 0.01, or if the p-value for the β parameter was less than or equal to 0.005. Additionally, in order to assess the effect of relaxing these criteria, a Type I error rate was developed using a less stringent criteria: $p_{\pi} \leq 0.04$; $p_{\beta} \leq 0.02$; and $p_{\alpha} \leq 0.04$. The Type I error rate, α, was the proportion of times a pair of datasets were deemed statistically different on a least one parameter allocating alpha as described.

For a sample size of 100, 471 (4.7%) of the 10,000 datasets were eliminated from consideration due to convergence and/or likelihood issues occurring for the initial parameter estimates or during generation of the standard errors for a dataset. Of the remaining 9,529 datasets, 321 were deemed statistically significant allocating alpha using the strict criteria described above for a Type I error rate of 3.4%. Using the less stringent criteria, the overall rate of rejecting the null hypothesis was 4.8% with 458 datasets deemed statistically different.

For a sample size of 200, 447 (4.5%) of the 10,000 datasets were eliminated from consideration due to convergence and/or likelihood issues occurring for the initial parameter estimates or during generation of the standard errors for a dataset. Of the remaining 9,553 datasets, 327 were deemed statistically significant allocating alpha using the strict criteria for a type I error rate of 3.4%. Using the less stringent criteria, the overall rate of rejecting the null hypothesis was 5.0% with 477 datasets deemed statistically different.

For a sample size of 500, 426 (4.3%) of the 10,000 datasets were eliminated from consideration due to convergence and/or likelihood issues occurring for the initial parameter

estimates or during generation of the standard errors for a dataset. Of the remaining 9,574 datasets, 356 were deemed statistically significant allocating alpha as described above for a Type I error rate of 3.7%. Using the less stringent criteria, the overall rate of rejecting the null hypothesis was 5.4% with 514 datasets deemed statistically different. The results for all sample sizes, using the two sets of criteria, are shown in Table 4.8.

Table 4.8 Assessment of Type I Error Rate

	Proportion of Datasets Deemed Statistically Different	
Sample Size	Criteria 1: $p_\alpha \leq 0.035$; $p_\pi \leq 0.01$; and $p_\beta \leq 0.005$	Criteria 2: $p_\alpha \leq 0.04$; $p_\pi \leq 0.04$; and $p_\beta \leq 0.02$
100	3.4%	4.8%
200	3.4%	5.0%
500	3.7%	5.4%

For all sample sizes, the Type I error rate was less than four percent using the stringent criteria. When using the less stringent criteria the Type I error rate still did not exceed five percent for sample sizes typically found in studies of interventions designed to increase adherences rates, suggesting it might be a useful technique for increasing the power of the test. Even at a sample size of 500, the observed Type I error rate using the less stringent criteria was just 5.4%.

Using the strict criteria, no systematic differences in the Type I error rates were noted within combinations of the parameters. For a sample size of 100, no combinations of β and α exceeded five percent when the mixing parameter π was 0.15. When π was 0.25, the combination of β and α both equal to 5 exhibited the largest Type I error rate of 6.1%. Two other combinations - β=3 and α=14, and β=11 and α=14 - exceeded five percent. However, the overall error when π was 0.25 was 3.4%. Results are shown in Table 4.9.

Table 4.9 Type I Error Rate per Parameter Combination: Strict Criteria, N=100

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	2.3% (87)	2.2% (136)	3.3% (182)	2.0% (197)	5.0% (200)
	5	3.5% (143)	3.7% (190)	4.0% (199)	1.5% (200)	4.5% (200)
	8	3.8% (186)	3.0% (199)	3.0% (200)	2.5% (200)	4.0% (200)
	11	4.1% (197)	3.5% (200)	3.5% (200)	3.5% (200)	3.5% (200)
	14	4.0% (199)	5.0% (200)	3.0% (200)	1.5% (200)	2.0% (200)
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	4.8% (104)	3.2% (158)	3.6% (196)	2.0% (200)	6.0% (200)
	5	2.5% (161)	6.1% (197)	3.0% (200)	3.0% (200)	1.0% (200)
	8	2.5% (198)	4.0% (200)	2.5% (200)	3.0% (200)	3.0% (200)
	11	3.0% (200)	4.0% (200)	4.5% (200)	5.0% (200)	5.5% (200)
	14	4.5% (200)	3.0% (200)	1.5% (200)	3.0% (200)	2.0% (200)

Note: 200 datasets were generated per parameter combination, however, only those that converged were used in analyses. Number in parentheses indicates the number of datasets that converged.

Results for a sample size of 200 are shown in Table 4.10. Results were similar, with four of the 50 parameter combinations (8%) of β and α exceeding five percent. However, none exceeded six percent and the overall Type I error rate for this sample size did not exceed four percent.

Table 4.10 Type I Error Rate per Parameter Combination: Strict Criteria, N=200

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	2.4% (82)	2.1% (146)	2.6% (194)	3.5% (199)	4.5% (198)
	5	2.9% (136)	3.1% (196)	2.5% (200)	2.0% (200)	4.0% (200)
	8	3.1% (191)	3.0% (200)	2.5% (200)	2.0% (200)	4.0% (200)
	11	2.5% (199)	4.5% (200)	3.0% (200)	3.0% (200)	3.0% (200)
	14	2.0% (200)	2.5% (200)	3.5% (200)	5.0% (200)	6.0% (200)
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	5.3% (76)	2.9% (170)	3.5% (200)	4.0% (200)	5.5% (200)
	5	3.6% (168)	2.0% (199)	3.0% (200)	3.0% (200)	3.0% (200)
	8	3.0% (199)	4.5% (200)	3.0% (200)	5.5% (200)	4.5% (200)
	11	4.0% (200)	2.5% (200)	5.0% (200)	3.5% (200)	4.5% (200)
	14	2.5% (200)	3.5% (200)	3.0% (200)	3.5% (200)	3.5% (200)

Note: 200 datasets were generated per parameter combination, however, only those that converged were used in analyses. Number in parentheses indicates the number of datasets that converged.

Results for a sample size of 500 are shown in Table 4.11. As might be expected Type I error rates rose at this sample size. Nine of the 50 parameter combinations (18%) of β and α exceeded five percent. However, only two of these exceeded six percent - one at 6.5% and one at 7.5% - while the overall Type I error rate for this sample size again did not exceed four percent.

These results indicate that the EMAAPE has adequate Type I error rates using the strict alpha control. For each sample size, the overall error rate was less than four percent while no combination of parameters exhibited error rates of concern. Those that exceeded five percent, usually did so by only one or two datasets.

Table 4.11 Type I Error Rate per Parameter Combination: Strict Criteria, N=500

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	2.9% (69)	1.4% (147)	5.0% (200)	3.5% (200)	4.0% (200)
	5	2.7% (149)	4.5% (199)	2.5% (200)	4.0% (200)	3.0% (200)
	8	3.0% (200)	2.5% (200)	2.0% (200)	3.5% (200)	6.0% (200)
	11	5.5% (200)	1.5% (200)	4.0% (200)	2.5% (200)	5.5% (200)
	14	1.5% (200)	3.5% (200)	1.5% (200)	4.0% (200)	5.0% (200)
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	2.0% (51)	1.1% (179)	6.0% (200)	6.5% (200)	5.0% (200)
	5	4.4% (180)	6.0% (200)	2.5% (200)	4.0% (200)	3.5% (200)
	8	4.0% (200)	2.0% (200)	2.5% (200)	5.5% (200)	5.0% (200)
	11	2.0% (200)	2.0% (200)	3.0% (200)	7.5% (200)	3.5% (200)
	14	4.5% (200)	3.5% (200)	4.0% (200)	5.5% (200)	2.5% (200)

Note: 200 datasets were generated per parameter combination, however, only those that converged were used in analyses. Number in parentheses indicates the number of datasets that converged.

Though error rates increased using the less stringent criteria, no systematic differences in the Type I error rates were noted within combinations of the parameters. For a sample size of 100, 19 of the 50 parameter combinations (38%) exceeded five percent, however none were exceptionally large (see Table 4.12).

Table 4.12 Type I Error Rate per Parameter Combination: Less Stringent Criteria, N=100

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	2.3% (87)	2.9% (136)	4.9% (182)	3.0% (197)	6.5% (200)
	5	3.5% (143)	4.2% (190)	5.5% (199)	3.0% (200)	5.0% (200)
	8	4.3% (186)	6.5% (199)	4.0% (200)	3.5% (200)	6.5% (200)
	11	5.1% (197)	4.5% (200)	4.5% (200)	4.0% (200)	4.5% (200)
	14	6.0% (199)	5.5% (200)	5.0% (200)	3.5% (200)	3.5% (200)
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	7.7% (104)	5.7% (158)	5.6% (196)	3.5% (200)	7.0% (200)
	5	4.3% (161)	6.6% (197)	5.5% (200)	5.5% (200)	3.0% (200)
	8	4.5% (198)	4.5% (200)	4.0% (200)	4.5% (200)	4.0% (200)
	11	6.5% (200)	5.0% (200)	6.0% (200)	6.5% (200)	6.0% (200)
	14	6.5% (200)	3.5% (200)	2.0% (200)	4.5% (200)	5.0% (200)

Note: 200 datasets were generated per parameter combination, however, only those that converged were used in analyses. Number in parentheses indicates the number of datasets that converged.

Results using the less stringent criteria for a sample size of 200 are shown in Table 4.13. Results were similar, with 21 of the 50 parameter combinations (42%) of β and α exceeding five percent. However, none exceeded eight percent and the overall Type I error rate for this sample size did not exceed five percent.

Table 4.13 Type I Error Rate per Parameter Combination: Less Stringent Criteria, N=200

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	2.4% (82)	2.1% (146)	5.7% (194)	3.5% (199)	7.6% (198)
	5	2.9% (136)	3.1% (196)	6.0% (200)	4.0% (200)	6.5% (200)
	8	3.7% (191)	6.0% (200)	3.0% (200)	4.0% (200)	5.0% (200)
	11	4.0% (199)	5.0% (200)	6.0% (200)	5.5% (200)	3.5% (200)
	14	4.5% (200)	4.5% (200)	6.5% (200)	6.5% (200)	7.5% (200)
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	6.6% (76)	2.9% (170)	5.0% (200)	6.5% (200)	8.0% (200)
	5	5.4% (168)	3.5% (199)	4.0% (200)	5.5% (200)	4.5% (200)
	8	3.5% (199)	6.0% (200)	5.0% (200)	6.5% (200)	6.5% (200)
	11	5.0% (200)	2.5% (200)	6.0% (200)	4.5% (200)	7.5% (200)
	14	3.5% (200)	6.0% (200)	5.0% (200)	4.0% (200)	5.0% (200)

Note: 200 datasets were generated per parameter combination, however, only those that converged were used in analyses. Number in parentheses indicates the number of datasets that converged.

Results for a sample size of 500 are shown in Table 4.14. As might be expected Type I error rates rose at this sample size, with the overall error rate exceeding five percent. Almost half of the 50 parameter combinations (48%) had error rates in excess of five percent with the largest exhibiting an error rate of 8.5% (π=0.25, β=11, α=11).

Table 4.14 Type I Error Rate per Parameter Combination: Strict Criteria, N=500

		α				
	β	**3**	**5**	**8**	**11**	**14**
π=0.15	**3**	2.9% (69)	2.0% (147)	6.0% (200)	5.0% (200)	6.5% (200)
	5	2.7% (149)	7.0% (199)	3.0% (200)	6.5% (200)	4.5% (200)
	8	5.0% (200)	3.0% (200)	4.5% (200)	4.0% (200)	7.5% (200)
	11	8.0% (200)	3.5% (200)	5.0% (200)	4.5% (200)	8.0% (200)
	14	2.5% (200)	7.5% (200)	3.0% (200)	5.5% (200)	7.5% (200)
		α				
	β	**3**	**5**	**8**	**11**	**14**
π =0.25	**3**	2.0% (51)	2.2% (179)	7.0% (200)	8.0% (200)	5.5% (200)
	5	5.0% (180)	8.0% (200)	4.5% (200)	5.0% (200)	5.0% (200)
	8	6.5% (200)	4.0% (200)	5.5% (200)	6.5% (200)	8.0% (200)
	11	3.5% (200)	4.0% (200)	6.5% (200)	8.5% (200)	6.0% (200)
	14	5.5% (200)	5.0% (200)	6.0% (200)	6.0% (200)	3.5% (200)

Note: 200 datasets were generated per parameter combination, however, only those that converged were used in analyses. Number in parentheses indicates the number of datasets that converged.

These results indicate that the EMAAPE has adequate Type I error rates using the less stringent criteria for sample sizes typically seen in studies of interventions to increase adherence rates. For sample sizes of 100 and 200, the overall error rate was less than five percent while no combination of parameters exhibited error rates of concern. Again, those that did exceed five percent usually did so by only one or two datasets. At larger sample sizes this criteria may not be sufficient given the elevated Type I error rates. However, this result is not overly concerning as studies of this magnitude are not typical, and if were to be conducted, would have sufficient power using the strict criteria for controlling the Type I error rate.

4.5 ASSESSMENT OF TYPE II ERROR

For each sample size, each of the 245,000 pairs of datasets discussed in section 3.3.3 were compared. In order to assess whether parameter estimates from two datasets created using different parameters were deemed statistically different, a t-score for the difference in the parameters was created using Equation 3.10 in section 3.5 above. P-values were then generated based on the t-statistic. Because it is possible that the same parameter may be used to generate portions of both curves, the distributions were deemed statistically significantly different if any of the three parameters was deemed significant. As above, in order to maintain an overall alpha level of 0.05, datasets were deemed statistically different if the p-value comparing the α parameters was less than or equal to 0.035, or if the p-value for the π parameter was less than or equal to 0.01, or if the β parameter was less than or equal to 0.005. Once again, in order to assess the effect of relaxing these criteria, we determined the Type II error rate using less stringent criteria: $p_{\pi} \leq 0.04$; $p_{\beta} \leq 0.02$; and $p_{\alpha} \leq 0.04$. As with the Type I error rate, the parameters and their standard errors were computed and compared to determine the proportion of times they were deemed different. The type II error rate, β, was the proportion of times that the algorithm failed to detect a statistical difference on a least one parameter. Additionally, the Sum of the Difference in Distributions was computed for each dataset compared.

4.5.1 Sum of the Differences in Distributions

For each of the 245,000 comparisons made for each sample size, the Sum of the Difference in Distributions (DD_SUM) was computed. The DD_SUM is the area between two curves and can

be thought of as the proportion of subjects positively affected by the intervention. Figure 4.3 displays an example of area between two curves for a hypothetical set of data.

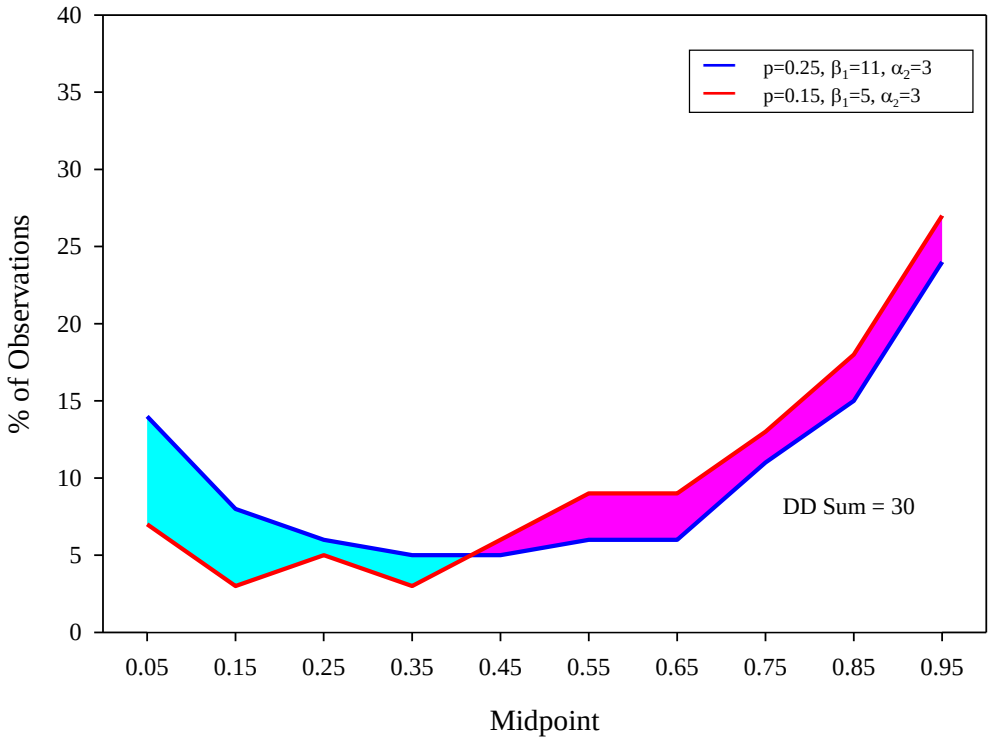


Figure 4.3 Hypothetical Comparison of Two Curves: π=0.25, β=11, α=3 v. π=15, β=5, α=3

The DD_SUM for the above example is 30 indicating that approximately 30% of the observations in one curve demonstrated higher levels or in terms of medication adherence that approximately 30% of the subjects in one intervention group exhibited better adherence than the subjects in the other. Table 4.15 shows the descriptive statistics for the DD_SUM for each sample size. The level of DD_SUM was divided into approximate tertiles representing small, medium, and large differences between datasets.

Table 4.15 Descriptive Statistics for the Sum of the Difference in Distributions

Sample Size	Mean	Range	% ≤10	% 10-20	% > 20	% of Datasets with Curves Crossing Once
100	17.8	2.0-124.0	36.9	32.5	30.6	3.9
200	18.6	1.0-120.0	36.9	30.7	32.4	7.7
500	19.5	0.4-106.8	36.3	27.9	35.9	13.5

Results across sample sizes were fairly consistent, though as the sample size increased the proportion of datasets with curves that crossed only once increased, indicating more well-defined distributions.

4.5.2 Type II Error Level

The results of analyses using the EMAAPE are shown in Table 4.16. For a sample size of 100, 19,666 (8.0%) of the 245,000 comparisons were excluded due to convergence and/or likelihood issues occurring for the initial parameter estimates or during generation of the standard errors for a dataset. Of the remaining 225,334 datasets, 147,783 (65.6%) were deemed statistically significant using the strict allocation of alpha as described above.

For a sample size of 200, 18,818 (7.7%) of the 245,000 comparisons were excluded due to convergence and/or likelihood issues occurring for the initial parameter estimates or during generation of the standard errors for a dataset. Of the remaining 226,182 datasets, 181,376 (80.2%) were deemed statistically significant using the strict allocation of alpha as described above.

For a sample size of 500, 19,308 (7.9%) of the 245,000 comparisons were excluded due to convergence and/or likelihood issues occurring for the initial parameter estimates or during generation of the standard errors for a dataset. Of the remaining 225,692 datasets, 204,385 (90.6%) were deemed statistically significant using the strict allocation of alpha as described above.

Table 4.16 P-values Obtained Using the EMAAPE by Sample Size Using Overall Alpha Level of 0.05

	p-value		
Sample Size	**Significant[1]**	**Marginally Significant[2]**	**Not Signficant[3]**
100	65.6%	6.1%	28.3%
200	80.2%	3.7%	16.1%
500	90.6%	1.4%	8.1%

1 - Defined as p-value$\leq$ 0.035 for α_1 v. α_2 or p-value$\leq$0.01 for ρ_1 v. ρ_2 or p-value$\leq$0.005 for β_1 v. β_2
2 - Defined as 0.035<p-value$\leq$ 0.07 for α_1 v. α_2 or 0.01<p-value$\leq$0.02 for ρ_1 v. ρ_2 or 0.005<p-value$\leq$0.01 for β_1 v. β_2
3 - Defined as p-value> 0.08 for α_1 v. α_2 or p-value>0.02 for ρ_1 v. ρ_2 or p-value>0.01 for β_1 v. β_2

At a sample size of 100 the EMAAPE perform marginally – 71.7% of all comparisons were at least marginally statistically significance level. However, only 65.6% were deemed statistically significant. At a sample size of 200 the EMAAPE performed better with just over 80% of comparisons exhibiting a statistically significant p-value, indicating a sample size of about 200 is optimal.

There were modest improvements in the power using the less stringent Type I error criteria, most notably for sample sizes of 100. The results are shown in Table 4.17.

Table 4.17 P-values Obtained Using the EMAAPE by Sample Size Using Less Stringent Criteria

	p-value		
Sample Size	**Significant[1]**	**Marginally Significant[2]**	**Not Signficant[3]**
100	70.7%	11.8%	17.5%
200	83.9%	7.0%	9.1%
500	92.4%	3.6%	4.0%

1 - Defined as p-value$\leq$ 0.04 for α_1 v. α_2 or p-value$\leq$0.04 for ρ_1 v. ρ_2 or p-value$\leq$0.02 for β_1 v. β_2
2 - Defined as 0.04<p-value$\leq$ 0.08 for α_1 v. α_2 or 0.04<p-value$\leq$0.08 for ρ_1 v. ρ_2 or 0.02<p-value$\leq$0.04 for β_1 v. β_2
3 - Defined as p-value> 0.08 for α_1 v. α_2 or p-value>0.08 for ρ_1 v. ρ_2 or p-value>0.04 for β_1 v. β_2

Furthermore, differences were observed by level of the sum of the difference in distributions. Table 4.18 displays the results of analyses by approximate tertiles of the DD_SUM.

Table 4.18 P-values Obtained Using the EMAAPE by Sample Size and Level of DD_SUM Using Strict Type I Error Criteria

	Value of Sum of the Difference in Distributions								
	≤ 10			10-20			>20		
	p-value			p-value			p-value		
Sample Size	Sign.[1]	Marg.[2]	N.S.[3]	Sign.[1]	Marg.[2]	N.S.[3]	Sign.[1]	Marg.[2]	N.S.[3]
100	63.4%	5.8%	30.8%	62.5%	6.6%	30.9%	71.6%	6.0%	22.4%
200	78.0%	4.0%	18.0%	78.3%	3.9%	17.7%	84.6%	3.1%	12.3%
500	90.1%	1.6%	8.3%	89.9%	1.1%	9.0%	91.5%	1.3%	7.1%

1 - Defined as p-value$\leq$ 0.035 for α_1 v. α_2 or p-value$\leq$0.01 for ρ_1 v. ρ_2 or p-value$\leq$0.005 for β_1 v. β_2
2 - Defined as 0.035<p-value$\leq$ 0.07 for α_1 v. α_2 or 0.01<p-value$\leq$0.02 for ρ_1 v. ρ_2 or 0.005<p-value$\leq$0.01 for β_1 v. β_2
3 - Defined as p-value> 0.08 for α_1 v. α_2 or p-value>0.02 for ρ_1 v. ρ_2 or p-value>0.01 for β_1 v. β_2

In general, the EMAAPE performed better for larger values of the DD_SUM (i.e., greater than 20). At a sample size of 100, the EMAAPE obtained statistical significance 71.6% of the time when DD_SUM was greater than 20. Results for DD_SUM less than 10 and between 10 and 20 were similar with just under two-thirds (63.4% and 62.5% respectively) reaching statistical significance.

Nearly 85% of the comparisons where DD_SUM was greater than 20 reached statistical significance for sample sizes of 200. Again, results for DD_SUM less than 10 and between 10 and 20 were similar with 78% of the dataset attaining statistical significance for both levels.

Results for samples sizes of 500 demonstrated even better results with at least 90% of the datasets reaching statistical significance regardless of the level of DD_SUM. The results across sample sizes indicate that the DD_SUM may act as an "effect size" and might be used to aid in determining sample sizes for potential studies.

Again, modest improvements were noted using the less stringent criteria for establishing statistical significance. Results are displayed in Table 4.19.

Table 4.19 P-values Obtained Using the EMAAPE by Sample Size and Level of DD_SUM Using Less Stringent Type I Error Criteria

	Value of Sum of the Difference in Distributions								
	≤ 10			10-20			>20		
	p-value			p-value			p-value		
Sample Size	**Sign.[1]**	**Marg.[2]**	**N.S.[3]**	**Sign.[1]**	**Marg.[2]**	**N.S.[3]**	**Sign.[1]**	**Marg.[2]**	**N.S.[3]**
100	66.6%	8.1%	25.3%	68.0%	15.1%	16.9%	78.8%	12.8%	8.3%
200	81.2%	4.6%	14.2%	82.3%	9.0%	8.6%	88.5%	8.0%	3.5%
500	92.2%	1.8%	6.0%	91.3%	4.0%	4.6%	93.5%	5.1%	1.4%

1 - Defined as p-value≤ 0.04 for α_1 v. α_2 or p-value≤0.04 for ρ_1 v. ρ_2 or p-value≤0.02 for β_1 v. β_2
2 - Defined as 0.04<p-value≤ 0.08 for α_1 v. α_2 or 0.04<p-value≤0.08 for ρ_1 v. ρ_2 or 0.02<p-value≤0.04 for β_1 v. β_2
3 - Defined as p-value> 0.08 for α_1 v. α_2 or p-value>0.08 for ρ_1 v. ρ_2 or p-value>0.04 for β_1 v. β_2

4.5.3 Comparison of Datasets Using Established Techniques

The result of analyses using established techniques are shown in Table 4.20. The EMAAPE is also displayed for comparison. At a sample size of 100, the t-test was clearly inadequate for distinguishing between datasets – slightly more than one-third (36.6%) of the comparisons were deemed statistically significant at a p-value of less than or equal to 0.05. Non-parametric tests, including the Wilcoxon Rank-Sum test and Kolmorgov-Smirnov (K-S) test, χ^2 test of association, and the 10-Level logistic regression performed better with slightly more than one-half of the comparisons deemed statistically significant. However, no established test performed better the EMAAPE.

The results for the sample size of 200 observations were similar. As expected, all of the tests performed better. Again, the t-test was the weakest test, with only 53.7% of the comparisons deemed statistically significant. Furthermore, none of the established tests performed as well as the EMAAPE, in which 80.2% of comparisons were significant.

Table 4.20 P-values Obtained Using Established Statistical Techniques and the EMAAPE by Sample Size

Sample Size=100			
	p-value		
Test	**Significant[1]**	**Marginally Significant[2]**	**Not Significant[3]**
T-test	36.6%	9.0%	54.4%
Wilcoxon Rank-Sum test	55.5%	6.7%	37.8%
χ^2 test	52.8%	6.8%	40.3%
K-S test	56.1%	7.1%	36.8%
5-Level Logistic Regression	46.8%	7.7%	45.4 %
10-Level Logistic Regression	53.8%	6.9%	39.4%
EMAAPE	65.6%	6.1%	28.3%
Sample Size=200			
	p-value		
Test	**Significant[1]**	**Marginally Significant[2]**	**Not Significant[3]**
T-test	53.7%	6.9%	39.4%
Wilcoxon Rank-Sum test	67.0%	5.2%	27.8%
χ^2 test	65.5%	5.2%	29.3%
K-S test	71.6%	6.9%	21.5%
5-Level Logistic Regression	61.0%	6.2%	32.8%
10-Level Logistic Regression	65.9%	5.5%	28.7%
EMAAPE	80.2%	3.7%	16.1%
Sample Size=500			
	p-value		
Test	**Significant[1]**	**Marginally Significant[2]**	**Not Significant[3]**
T-test	69.5%	4.5%	26.0%
Wilcoxon Rank-Sum test	79.5%	3.2%	17.3%
χ^2 test	77.5%	3.2%	19.3%
K-S test	91.1%	1.9%	7.0%
5-Level Logistic Regression	75.0%	3.5%	21.5%
10-Level Logistic Regression	78.4%	3.2%	18.4%
EMAAPE	90.6%	1.4%	8.1%

NOTE: Only datasets that converged using the EMAAPE were evaluated (n=225,334 for a sample size of 100; n=226,182 for a sample size of 200; and n=225,692 for a sample size of 500).

1 - Defined as p-value$\leq$0.05 for established tests and p-value$\leq$ 0.035 for α_1 v. α_2 or p-value$\leq$0.01 for ρ_1 v. ρ_2 or p-value$\leq$0.005 for β_1 v. β_2 for the EMAAPE

2 - Defined as 0.05<p-value$\leq$0.10 for established tests and 0.035<p-value$\leq$ 0.07 for α_1 v. α_2 or 0.01<p-value$\leq$0.02 for ρ_1 v. ρ_2 or 0.005<p-value$\leq$0.01 for β_1 v. β_2 for otherwise non-significant results for the EMAAPE

3 - Defined as p-value>0.10 for established tests and p-value> 0.07 for α_1 v. α_2 and p-value>0.02 for ρ_1 v. ρ_2 and p-value>0.01 for β_1 v. β_2 for the EMAAPE

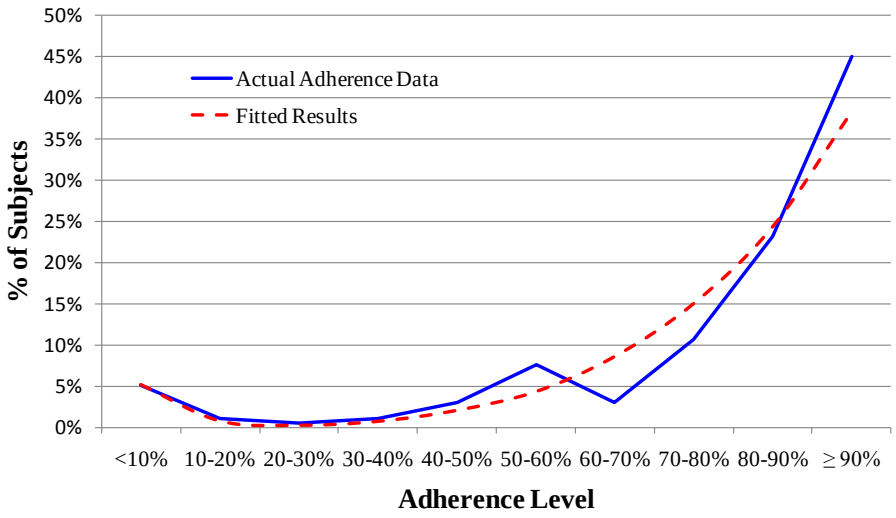
Results for samples of size 500 were again better for all tests. At this sample size most of the established tests approached 80% power with the exception of the K-S test which demonstrated results similar to the EMAAPE. While this test performed adequately, of the two, only the EMAAPE provided parameter estimates describing the shape of the distribution.

4.6 ANALYSIS OF THE RHEUMATOID ARTHRITIS STUDY DATA

4.6.1 Baseline Data

To evaluate the performance of the EMAAPE to produce parameter estimates, baseline data from the second Rheumatoid Arthritis Study was analyzed. A total of 198 subjects presented for the RA Study and provided data during the first six weeks of the study. One in twenty subjects demonstrated very poor adherence rates (<10%) while more than two-thirds (68.2%) demonstrated good adherence rates (>80%). As seen in Figure 4.4 the distribution was J-shaped in nature.

The EMAAPE was used to estimate the parameters of the distribution and the fitted line is also shown in Figure 4.4. The EMAAPE converged after 16 iterations, taking less than 10 seconds of real time. The parameter estimates were π=0.061; β=18.4; and α=5.0. Visual inspection of the fitted line shows that it is a good representation of the adherence distribution observed in this study.



Fitted Distribution Based on Mixed Beta Model: $0.061\beta(18.4,1)+(1-0.061)\beta(1,5.0)$

Figure 4.4 Observed and Fitted Baseline Adherence Distributions for the Rheumatoid Arthritis Study

4.6.2 Six Month Data

In order to evaluate the EMAAPE for comparing the two treatment groups from the RA studies, a total of 220 subjects were evaluated. The Usual Care (UC) group consisted of 93 subjects, while the Telephone Intervention (TI) group consisted of 127 subjects. The EMAAPE was used to compare these two groups, using medication adherence for the 28 days leading up to the midpoint of the treatment phase of the studies (6 months post enrollment). Figure 4.5 displays the distributions for the two treatment groups. The mean adherence rate for the TI group was 65.8% (median=71.4%), while the mean adherence rate for the UC group was 60.4% (median=75.0%). Additionally, a larger proportion of the UC group demonstrated adherences levels at or above 80% (39.8% v. 36.2%, p=0.59). The DD_SUM for these two groups was 18, indicating that approximately 18% of the patients in the TI group demonstrated better adherence rates than those in the UC group.

Using a parametric t-test the two groups were not statistically significantly different (t=1.3, p=0.1975). Using the EMAAPE the parameters for the TI group were π=0.17 (SE=0.035); β=2.8 (SE=0.51); and α=2.9 (SE=0.35). The parameters for the UC group were π=0.34 (SE=0.081); β=3.6 (SE=1.42); and α=4.1 (SE=1.01). It is apparent here that for each parameter that the variances are not equal, as each ratio exceeds 3.8[5]. Observed and fitted results are shown in Figure 4.5.

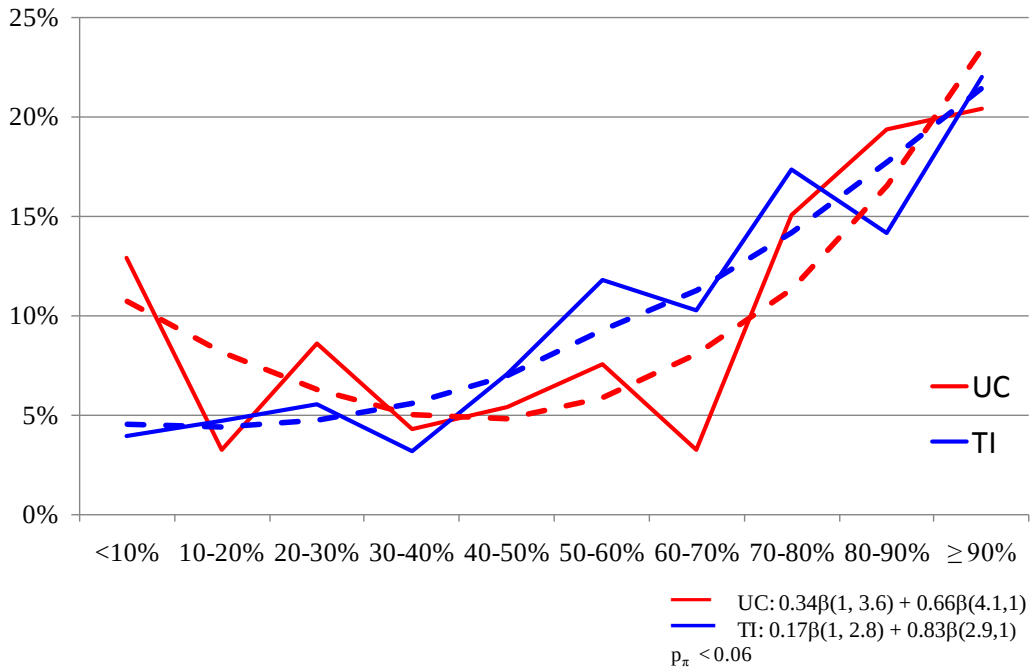


Figure 4.5 Observed and Fitted RA Study Adherence Distributions Comparing Usual Care to Telephone Intervention: Six Months Post Intervention

For the UC group the EMAAPE converged after eight iterations. However, several of the jackknifed datasets used to compute the standard errors did not converge. The problem was rectified by systematically using several different initial starting points for the EMAAPE and selecting the results yielding the greatest maximum likelihood estimate from those that

[5] p-value for F ratio <0.0001 for F=3.8, df_1=92, df_2=126

converged. The EMAAPE converged after 13 iterations for the TI group and jackknifed datasets converged without incident.

Based on the difference in the mixing parameter π, final p value was 0.060 indicating the two groups approached statistical significance. As the curve for the TI treatment group had fewer subjects in the left hand component and more subjects in the right had component (based upon the difference between the mixing proportions, π, and visual inspection) it was deemed that the treatment was more successful in raising medication adherence levels of the subjects receiving it.

This result was not observed based on the results of the parametric t-test. Results similar to those of the t-test were obtained using non-parametric techniques. Table 4.21 displays the p-values for all statistical tests conducted.

Table 4.21 Results of Comparisons Between the Intervention and Usual Care Groups of the Rheumatoid Arthritis Data

Test	p-value
T-test	0.1975
Wilcoxon Rank-Sum test	0.4527
χ^2 test	0.5900
K-S test	0.3849
5-Level Logistic Regression[1]	0.5071
10-Level Logistic Regression[1]	0.4378
EMAAPE[2]	0.0597

1 - Wald Test
2 - Based on difference in the mixing parameter, π

5.0 DISCUSSION

The Expectation-Maximization Algorithm has been used to estimate parameters of distributions when one or more parameters are impossible to estimate due to constraints of the distribution. Mixture distributions are a prime example of distributions with constraints as the mixing parameter usually cannot be estimated in closed form. We have demonstrated that adherence data, J-shaped in nature, can be approximated as the mixture of two beta distributions. The EMAAPE provides a technique to estimate the parameters of this mixed distribution. Additionally, it provides a technique for comparing two adherence distributions through the use of the parameter estimates and their associated standard errors.

Analysis of adherence data provides a challenge because of the form of its distribution. Transforming the data is a common technique used to handle this challenge, however, several different methods, including Box-Cox transformations, were attempted using simulated data and none were successful at inducing normality. The use of parametric techniques on simulated data in violation of the normality assumption were also shown to be problematic in that only slightly more than one in three times were the two distributions deemed statistically different for sample sizes of 100 per group and only slightly more than one in two for sample sizes of 200 per group indicating these tests lacked sufficient power to detect differences when assessing J-shaped distributions. Only when sample sizes approached 500 per group were results acceptable.

However, sample sizes of this magnitude are often not feasible for studies of interventions designed to increase medication adherence.

Other non-parametric statistical techniques were assessed for the analysis of simulated adherence data. These tests included the Wilcoxon Rank-Sum test, χ^2 Tests of Association of dichotomized data, the Kolmogorov-Smirnov test, and 5- and 10-Level Logistic Regression. These techniques performed better than the parametric t-test, but still performed inadequately. For sample sizes of 100 per group, only about one in two were statistically significant, while this number increased to about two of three for samples of 200 per group. Again, results for these tests were better at sample sizes of 500 or more per group but as noted above studies of this size are often not feasible.

Of these non-parametric tests considered the Kolmogorov-Smirnov test performed the best at every sample size in simulations among the established tests considered. However, one inherent flaw is that it fails to provide any information regarding the shape of those curves, information valuable to researchers interested in determining the full impact of interventions.

These results clearly show the need for the development of a statistical technique that accounts for the J-shape distribution of adherence data. Current techniques of analysis are not adequate for properly evaluating differences between interventions aimed at improving medication adherence. Therefore, an intervention that might result in improved adherence among patients may be overlooked, with the result being the loss of a tool that could decrease morbidity, mortality, and health care costs. The use of the EMAAPE to estimate parameters of a mixed beta distribution is such a technique. The results showed that using a composite beta distribution, a J-shaped distribution was reasonably simulated. It is also clear by the nature of

beta distribution that the values of α and β provide information regarding the levels of observed adherence. Larger values of α and β, relative to smaller ones, result in steeper curves translating into an increased number of observations for that portion of the curve. It was shown via simulations that for data generated using a composite beta distribution the proportion of observations at or above 0.8 (an indicator of good adherence) increased with increasing values of the α parameter, indicating its potential use for interpreting parameters estimated using the EMAAPE. This phenomenon was not noted with the β parameter, likely due to the lower proportion of observations generated for the left hand portion of the curve. Additionally, the mixing parameter π serves as an indicator of better v. worse adherence with larger values indicative of lower adherence as more data points reside in the lower portion of the curve. Thus, all three parameters provide information regarding the nature of the distribution and must be interpreted simultaneously.

In simulations the EMAAPE was shown to be able to produce parameter estimates of the composite beta distribution with little or no bias across a range of sample sizes. Convergence problems and diverging maximum likelihood estimates were minimal. Furthermore, the EMAAPE produced standard error estimates providing a mechanism for distinguishing curves from two different models. At smaller samples sizes, those typically seen in studies of medication adherence, the EMAAPE performed better than established techniques. Two-thirds of the datasets compared at sample sizes of 100 per group were deemed statistically different compared to one-third to one-half of those using established tests. At a sample size of 200, about 80% were deemed statistically different compared to a little more than half using t-tests and about two-thirds using non-parametric tests. Thus, an optimal sample size when considering this new technique appears to be approximately 200 subjects per group. Additionally, we

demonstrated that the power of the EMAAPE could be improved for smaller samples by using less stringent criteria on the overall alpha level without unduly inflating the Type I error rate.

The EMAAPE appeared to provide reasonable parameter estimates when applied to data collected from the RA study. While some parameter estimates were outside the range used in simulations, fitted curves appeared to adequately depict the data. Furthermore, standard errors generated for the parameters estimated were reasonable indicating the estimates were relatively stable. One potential problem with actual adherence data is the clustering of data points due to a finite number of possible denominators used in adherence calculations. For example, for a two week period, there are only 15 possible data points. However, this problem can be addressed by increasing the number of days over which the adherence measure is calculated. For this study we used 28 days after initial attempts using 21 days demonstrated convergence problems. Another possible solution would be to increase the complexity of the adherence measure by accounting for dosing intervals which would provide a measure of the proportion of time a subject is receiving an adequate therapeutic effect, increasing the potential number of unique data points. While more complex adherence calculations would result in decreased adherence rates (Rohay, Dunbar-Jacob, Sereika, & Kwoh, 1996), the method of calculation would be applied across all treatment groups and thus would not impart any undue bias.

We also demonstrated that the EMAAPE could be used to examine differences between the adherence distributions of the two groups of patients from the RA study. Based on our strict control of the overall alpha level for the three comparisons our results only approached statistical significance (p_{π}-value$\leq$ 0.06). However, results from the simulations across sample sizes showed that using criteria of $\alpha_{\pi} \leq 0.04$, $\alpha_{\beta} \leq 0.02$, and $\alpha_{\alpha} \leq 0.04$ still resulted in less than five

percent of comparisons leading to an incorrect rejection of the null hypothesis that the two groups differed statistically. Using these criteria the results of the RA study more strongly suggested that the Telephone Intervention was successful in improving adherences rates as a smaller percentage of subjects were in the lower portion of the curve (17% v. 34%).

Furthermore, examination of the data from this study pointed out the inherent difficulty of interpreting results based on established techniques. Relying on measures of central tendency, such as the mean and median, and/or dichotomizing the data resulted in conflicting and ultimately confusing conclusions. Based on the means, one could conclude that the Telephone Intervention was superior to Usual Care, while the opposite conclusion would be drawn based on the medians or dichotomizing the data. Using the EMAAPE to estimate the parameters of the distribution provided a clearer picture of the results. As described above both the mixing parameter π and the parameter α are directly interpretable for describing differences between treatment groups. Therefore, the results using the EMAAPE, based specifically on the mixing parameter π, more clearly demonstrate that the Telephone Intervention group attained superior adherence rates.

As mentioned above, the EMAAPE can be difficult to implement. Convergence using data collected in real study settings may become problematic as the number of unique data points may be inadequate. Additionally, while convergence problems were minimal in simulations, they did occur, as did decreasing maximum likelihoods. The use of different initial values for the EMAAPE is one strategy to deal with these problems, though others need to be explored. These problems were mainly observed in a small number of parameter values tested, however, additional values should be examined to better understand the full range of parameter values

leading to J-shaped distributions. Additionally, examination of datasets with fewer observations, in the range of 50 observations per group, should be conducted to assess the utility of the EMAAPE in smaller studies.

Finally, examination of the interrelationship between the three parameters of the mixed beta model was beyond the scope of this project. For this project two arbitrary selections of alpha levels were used to determine statistically significantly different models. The first was designed to ensure an absolute overall alpha level of no more than five percent (strict criteria) while the second was based on simulation results demonstrating a Type I error rate of less than five percent (less stringent criteria). While the less stringent criteria resulted in overall Type I error rates less than five percent for smaller sample sizes (those more likely to be seen in studies of interventions to increase adherence rates), some combinations of parameters exhibited error rates exceeding five percent. There appeared to be nothing systematic about those combinations, though, and their error rates were not exceedingly high. However, other less stringent criteria need to be explored that may increase power while maintaining adequate Type I error rates for every combination of the parameters. Therefore, understanding the interrelationship between the three parameters is an important step in determining the optimal criteria for determining statistical significance.

In conclusion, estimation of the parameters of a composite beta distribution appears to be a favorable solution to characterize adherence data and can provide a sound statistical technique for comparing adherence data across multiple groups. The EMAAPE provided a more powerful test while not demonstrating an inflated alpha level, even when using less stringent criteria to account for the multiple comparisons being made. Use of statistical analyses of this nature will

impact public health by helping researchers in this area better understand the nature of this data and hopefully better assess interventions designed to improve medication adherence leading to decreased patient morbidity and mortality, and decreased health care costs.

APPENDIX A

EMAAPE SAS MACRO

```
%macro revised_em(orig_ds=, p_hat=0.04, b_hat=3, a_hat=3, N=, tol=, max_its=999, final_DS=,
emtime_ds=);


data TIME_START_EM;
     START_D=DATE();
     START_T=TIME();
run;

PROC TRANSPOSE DATA=&orig_ds PREFIX=X OUT=TRAW;
      by PS DS REPLICATE;
      VAR X;
run;

data rtheta;
     set traw;
     ITERATION=0;
     p_hat=&p_hat;
     b_hat=&b_hat;
     a_hat=&a_hat;
     SUM_LIKE=0;

     keep ps ds replicate iteration p_hat b_hat a_hat SUM_LIKE;
run;

data EM;
     set traw;
     ITERATION=1;

     ARRAY X    (*) X1-X&N;
     ARRAY LX1X (*) LOG_1_X1-LOG_1_X&N;
     ARRAY LXX  (*) LOG_X1-LOG_X&N;
     ARRAY PI1X (*) PI1X1-PI1X&N;
     ARRAY PI2X (*) PI2X1-PI2X&N;
     ARRAY P1NX (*) P1NX1-P1NX&N;
     ARRAY P2NX (*) P2NX1-P2NX&N;
     ARRAY P1LX (*) P1LX1-P1LX&N;
     ARRAY P2LX (*) P2LX1-P2LX&N;
     ARRAY Z    (*) Z1-Z&N;
     ARRAY L1   (*) L1X1-L1X&N;
     ARRAY L2   (*) L2X1-L2X&N;
     ARRAY LIKE (*) LIKE1-LIKE&N;

     SUM_P1_LOG_X=0;
     SUM_P2_LOG_X=0;
     SUM_P1N=0;
     SUM_P2N=0;
     SUM_LIKE=0;
     SUM_Z=0;

     do j=1 to &n;
        LX1X[j]=log(1-x[j]);
        LXX[j]=log(x[j]);
        PI1X[j]=&p_hat*PDF('BETA',x[j],1,&b_hat,0,1);
        /* Get Probability of x from Left BETA DIST  */
        PI2X[j]=(1-&p_hat)*PDF('BETA',x[j],&a_hat,1,0,1);
        /* Get Probability of x from Right BETA DIST */
        P1NX[j]=pi1x[j]/(pi1x[j]+pi2x[j]);
        /* Get Probability of pi1 from total prob    */
        P2NX[j]=pi2x[j]/(pi1x[j]+pi2x[j]);
        /* Get Probability of pi2 from total prob    */
```

```
        P1LX[j]=p1nx[j]*LX1X[j];
        P2LX[j]=p2nx[j]*LXX[j];

        if p1nx[j]>p2nx[j] then z[j]=1; else z[j]=0;

        L1[j]=(log(p1nx[j])+max(log(PDF('BETA',x[j],1,&b_hat,0,1)),-705));
        L2[j]=(log(p2nx[j])+max(log(PDF('BETA',x[j],&a_hat,1,0,1)),-705));
        LIKE[j]=p1nx[j]*L1[j]+p2nx[j]*L2[j];

        SUM_P1_LOG_X=SUM_P1_LOG_X+P1LX[j];
        SUM_P2_LOG_X=SUM_P2_LOG_X+P2LX[j];
        SUM_P1N=SUM_P1N+P1NX[j];
        SUM_P2N=SUM_P2N+P2NX[j];
        SUM_LIKE=SUM_LIKE+LIKE[j];
        SUM_Z=SUM_Z+Z[j];
     end;

     p_hat=SUM_P1N/&n;
     b_hat=-SUM_P1N/SUM_P1_LOG_X;
     a_hat=-SUM_P2N/SUM_P2_LOG_X;

run;

data theta_i;
     set em;
     FIRST_LIKE=SUM_LIKE;
     keep ps ds replicate iteration p_hat a_hat b_hat SUM_LIKE FIRST_LIKE;
run;

data rtheta;
     set rtheta theta_i;
run;

proc sort data=rtheta; by ps ds replicate iteration; run;

data rtheta;
     set rtheta;
     DIF_LIKE=SUM_LIKE-lag(SUM_LIKE);
     ADIF_LIKE=abs(SUM_LIKE-lag(SUM_LIKE));
     if iteration=0 then DIF_LIKE=9999;
     if iteration=0 then ADIF_LIKE=9999;
run;

%let i=1;                 /* Initialize iteration counter                */
%do %while (&i<&max_its); /* While iterations are less than max          */
    %let i=%eval(&i+1);   /* Increment iteration counter                 */
    DM LOG 'clear' continue;            /* Clear log window */

    data em2;
         merge em theta_i(in=A); by ps ds replicate;

         if A;
         ITERATION=&i;

         ARRAY X    (*) X1-X&N;
         ARRAY LX1X (*) LOG_1_X1-LOG_1_X&N;
         ARRAY LXX  (*) LOG_X1-LOG_X&N;
         ARRAY PI1X (*) PI1X1-PI1X&N;
         ARRAY PI2X (*) PI2X1-PI2X&N;
         ARRAY P1NX (*) P1NX1-P1NX&N;
         ARRAY P2NX (*) P2NX1-P2NX&N;
         ARRAY P1LX (*) P1LX1-P1LX&N;
         ARRAY P2LX (*) P2LX1-P2LX&N;
         ARRAY Z    (*) Z1-Z&N;
         ARRAY L1   (*) L1X1-L1X&N;
         ARRAY L2   (*) L2X1-L2X&N;
         ARRAY LIKE (*) LIKE1-LIKE&N;

         SUM_P1_LOG_X=0;
         SUM_P2_LOG_X=0;
         SUM_P1N=0;
```

```
     SUM_P2N=0;
     SUM_LIKE=0;
     SUM_Z=0;


     do j=1 to &n;
        LX1X[j]=log(1-x[j]);
        LXX[j]=log(x[j]);
        PI1X[j]=p_hat*PDF('BETA',x[j],1,b_hat,0,1);
        /* Get Probability of x from Left BETA DIST  */
        PI2X[j]=(1-p_hat)*PDF('BETA',x[j],a_hat,1,0,1);
        /* Get Probability of x from Right BETA DIST */

        P1NX[j]=pi1x[j]/(pi1x[j]+pi2x[j]);
        /* Get Probability of pi1 from total prob    */
        P2NX[j]=pi2x[j]/(pi1x[j]+pi2x[j]);
        /* Get Probability of pi2 from total prob    */
        P1LX[j]=p1nx[j]*LX1X[j];
        P2LX[j]=p2nx[j]*LXX[j];

        if p1nx[j]>p2nx[j] then z[j]=1; else z[j]=0;

        L1[j]=(max(log(p1nx[j]),0)+max(log(PDF('BETA',x[j],1,b_hat,0,1)),-705));
        L2[j]=(max(log(p2nx[j]),0)+max(log(PDF('BETA',x[j],a_hat,1,0,1)),-705));
        LIKE[j]=p1nx[j]*L1[j]+p2nx[j]*L2[j];

        SUM_P1_LOG_X=SUM_P1_LOG_X+P1LX[j];
        SUM_P2_LOG_X=SUM_P2_LOG_X+P2LX[j];
        SUM_P1N=SUM_P1N+P1NX[j];
        SUM_P2N=SUM_P2N+P2NX[j];
        SUM_LIKE=SUM_LIKE+LIKE[j];
        SUM_Z=SUM_Z+Z[j];

     end;

     p_hat=SUM_P1N/&n;
     b_hat=-SUM_P1N/SUM_P1_LOG_X;
     a_hat=-SUM_P2N/SUM_P2_LOG_X;

run;

proc contents data=em2 out=e NOPRINT;  run;

data empty;
     set e;
     if _N_=1;
     keep nobs;
    CALL SYMPUT('NOBS', nobs);  /* Create macro variable for difference        */
run;

%if &NOBS=0 %then %GOTO EXIT;

data theta_i_1;
     set theta_i;
run;


data theta_i;
     set em2;
     keep ps ds replicate iteration p_hat a_hat b_hat SUM_LIKE;
run;

data theta_dif;
     set theta_i_1 theta_i;
run;

proc sort data=theta_dif; by ps ds replicate iteration; run;

data theta_dif1;
     set theta_dif; by ps ds ;
     FIRST_LIKE=lag(FIRST_LIKE);
```

```
          DIF_LIKE=SUM_LIKE-lag(SUM_LIKE);
          ADIF_LIKE=abs(SUM_LIKE-lag(SUM_LIKE));

          if first.DS then DIF_LIKE=.;
          if first.DS then ADIF_LIKE=.;
     run;
     data theta_i;
          set theta_dif1; by ps ds replicate iteration;
          if last.replicate;
     run;

     data rtheta;
          set rtheta theta_i;
     run;
     data theta_i;
          set theta_dif1; by ps ds replicate iteration;
          if last.replicate;
          if adif_like<&tol then delete;
     run;
     %put E-M Iteration # &i;

%end;

%EXIT:           /* If difference criteria met then exit   */

proc sort data=rtheta; by PS ds replicate iteration; run;

data %str(&FINAL_DS)_ITERATIONS;
     set rtheta; by PS ds replicate iteration;
     if ds>0;
     p=input(substr(PS,8,2), f3.0);  /* Initialize p1 (% of observations from left)  */
     b=input(substr(PS,11,2), f3.0); /* Initialize p1 (% of observations from left)  */
     a=input(substr(PS,14,2), f3.0); /* Initialize p1 (% of observations from left)  */
run;

data &FINAL_DS;
     set rtheta; by PS ds replicate iteration;
     if last.replicate;
     if ds>0;
     p=input(substr(PS,8,2), f3.0);  /* Initialize p1 (% of observations from left)  */
     b=input(substr(PS,11,2), f3.0); /* Initialize p1 (% of observations from left)  */
     a=input(substr(PS,14,2), f3.0); /* Initialize p1 (% of observations from left)  */
run;

data TIME_END;
     END_D=DATE();
     END_T=TIME();
run;

data &emtime_ds;
     merge time_start_EM time_end;
run;

proc print data=&emtime_ds;
     format START_D END_D mmddyy. start_t end_t time.;
run;
%mend revised_em;
```

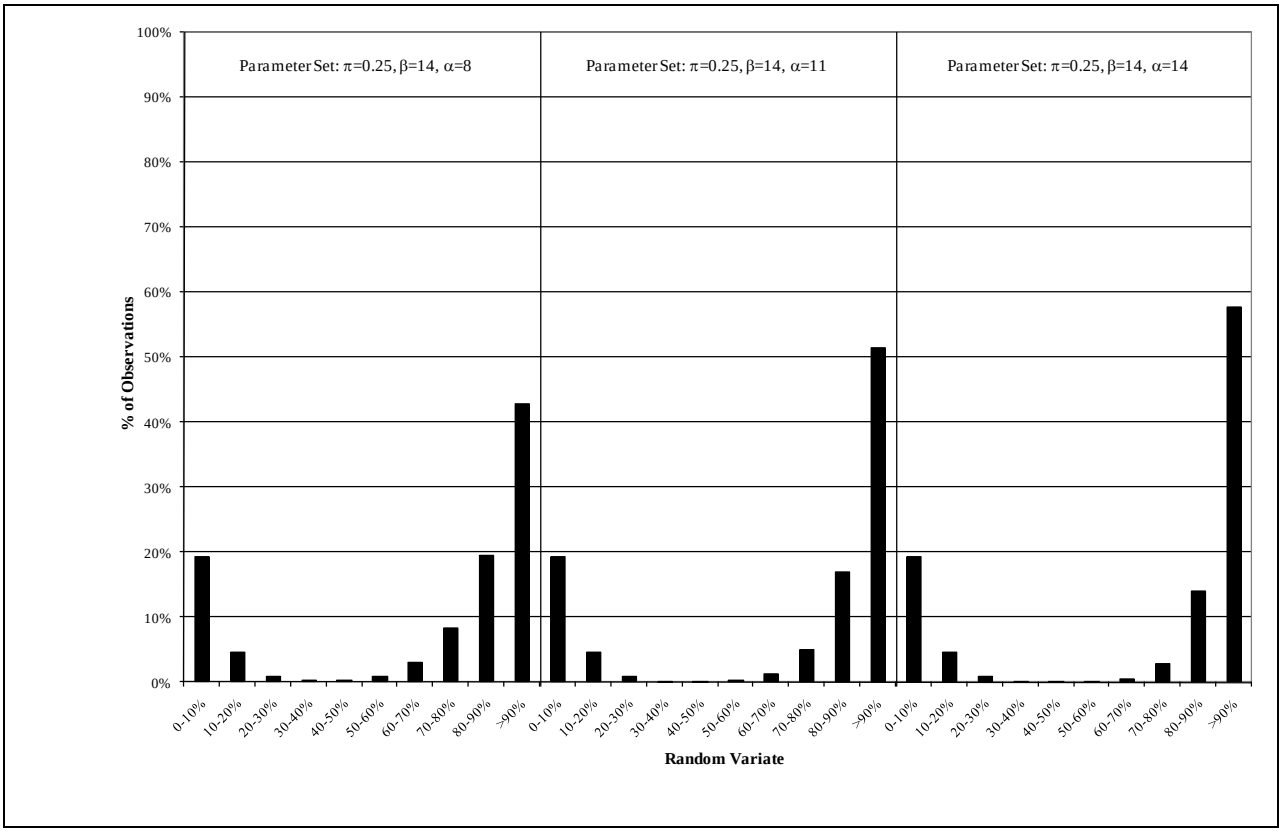
APPENDIX B

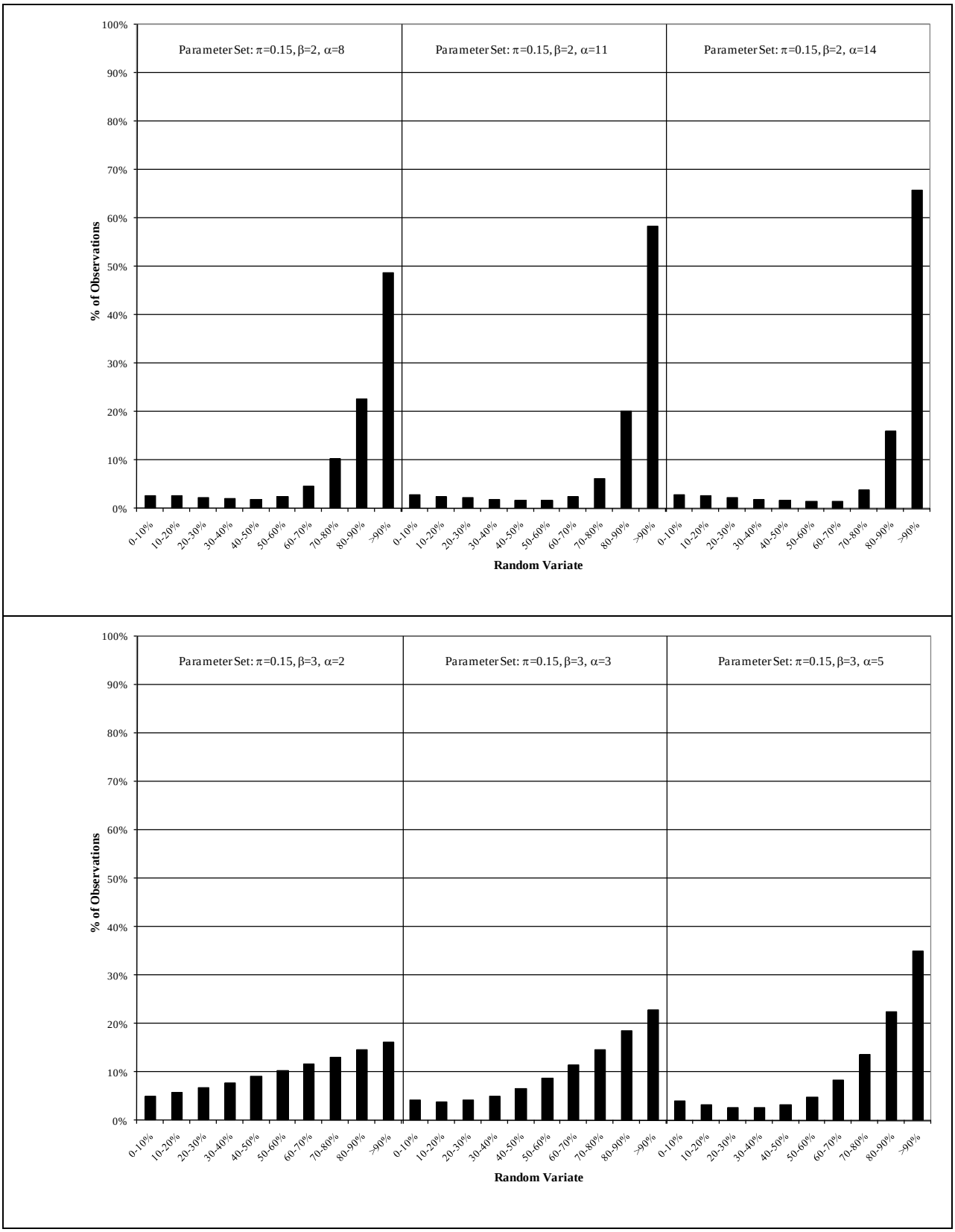
SAS MACRO FOR COMPUTING STANDARD ERROR

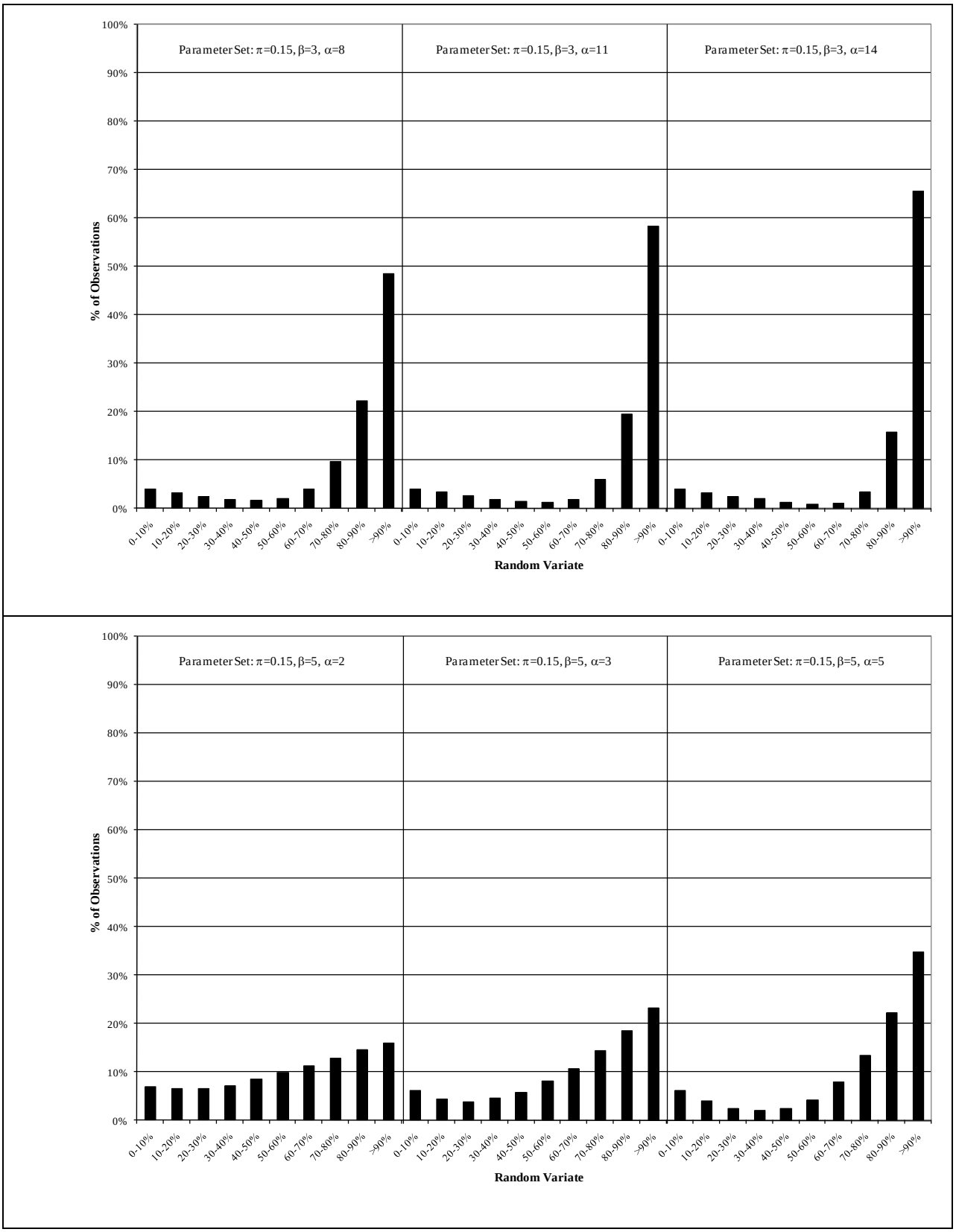
```
%macro jack_variance(ds=, SS=, final=, max_its=1000);
proc means data=&ds NOPRINT;
     by ps ds;
     var p_hat b_hat a_hat;
     output out=EST1 var=p_hat_j b_hat_j a_hat_j;
     /* NOTE: VAR=(x-x_bar)^2/(n-1) */
     where iteration<&max_its & FIRST_LIKE<SUM_LIKE;
run;

data &final;
     set est1(rename=(_FREQ_=JACK_SS));

     array d p_hat_j b_hat_j a_hat_j;
     array v v_p_hat_JACK v_b_hat_JACK v_a_hat_JACK;
     array s sd_p_hat_JACK sd_b_hat_JACK sd_a_hat_JACK;
     do over d;
        v=((JACK_SS-1)**2*d)/JACK_SS;
         /* NOTE: VAR_j=[(N-1)(x-x_bar)^2]/n so Mult d by [N-1]^2 */
        s=sqrt(v);
     end;
     drop p_hat_j b_hat_j a_hat_j;
run;
/* NOTE: See SE ESTIMATES QA.XLS for check of formula */

%mend jack_variance;
```

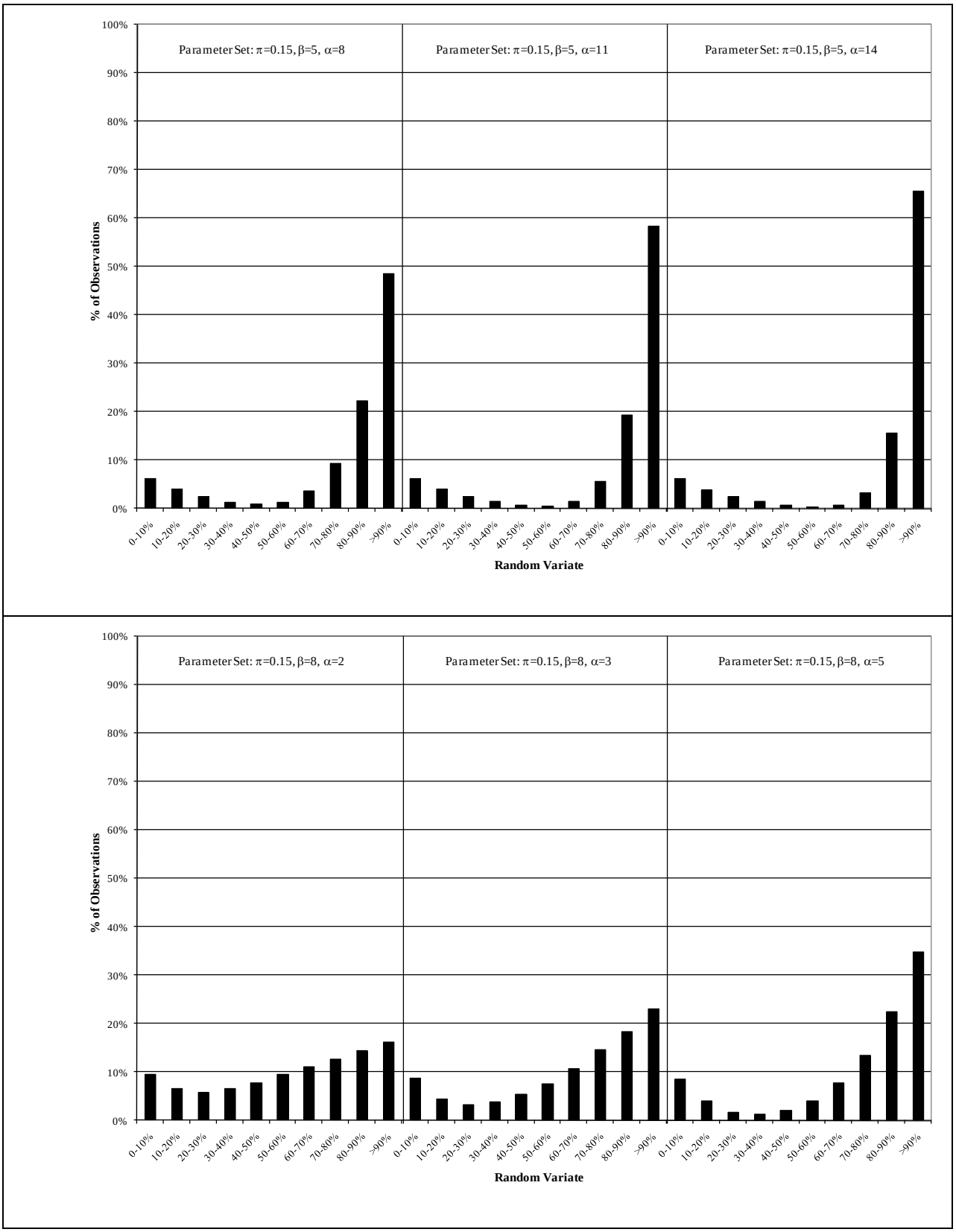
APPENDIX C

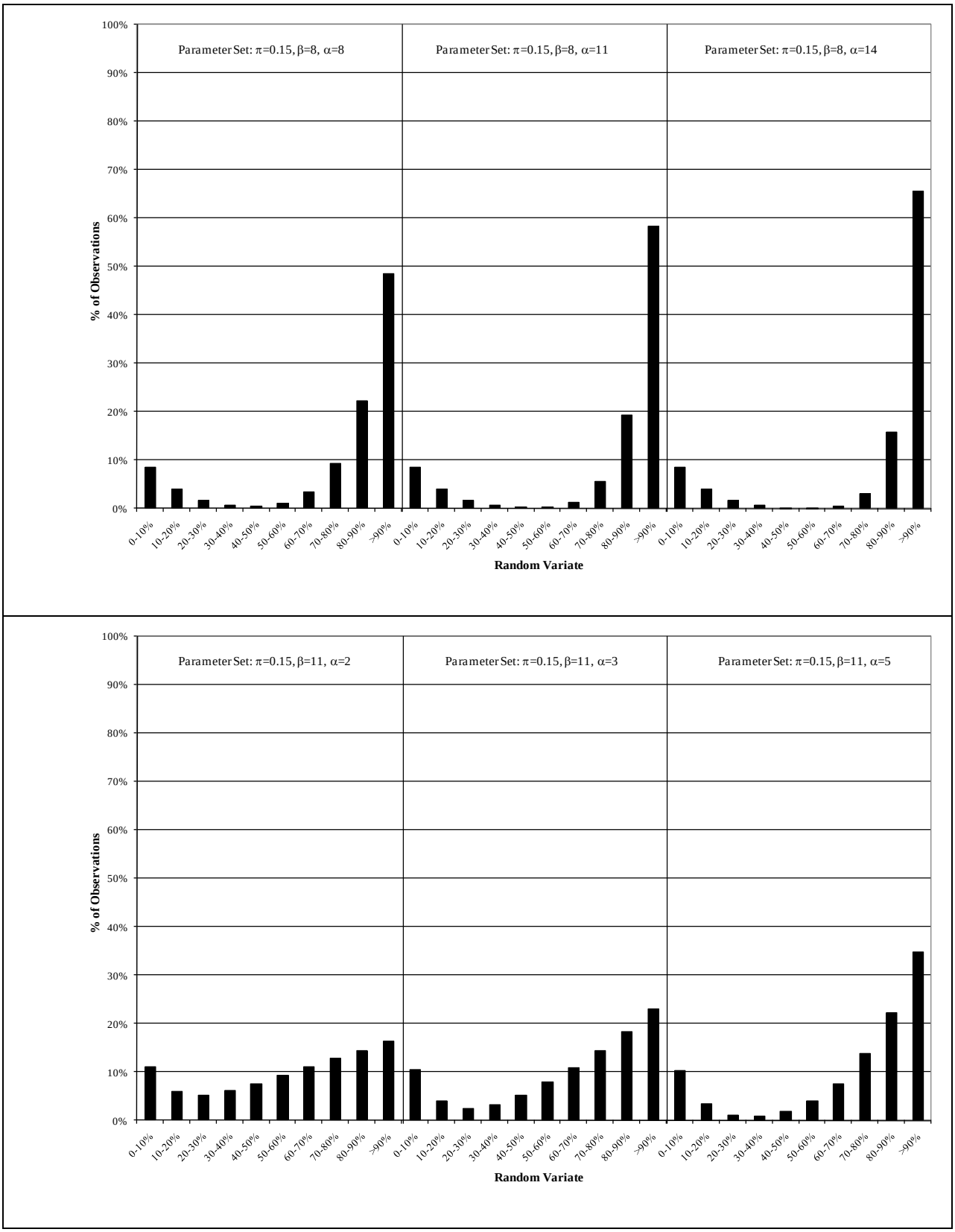
HISTOGRAM OF 100,000 RANDOM DEVIATES FROM ALL MIXED BETA DISTRIBUTIONS

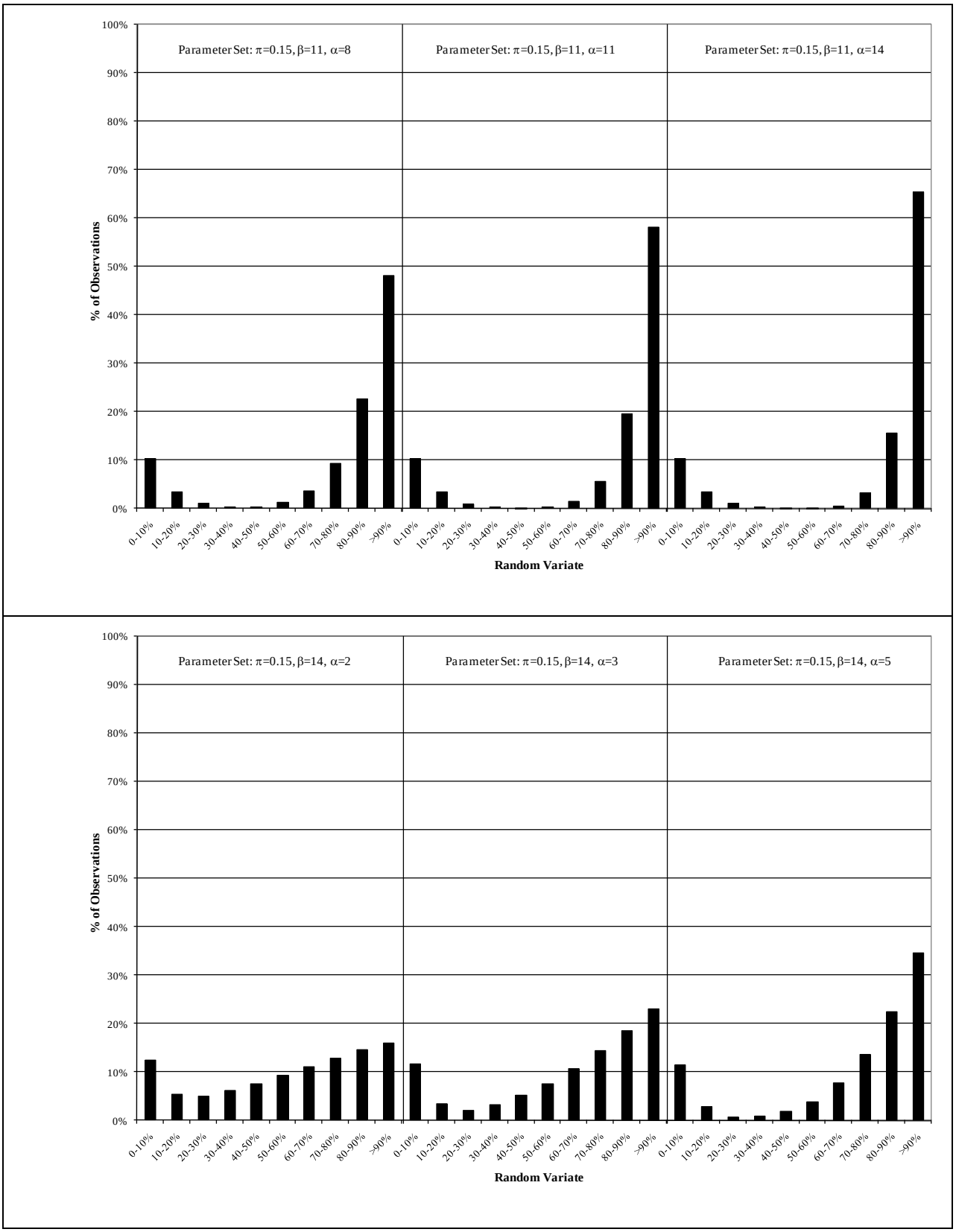


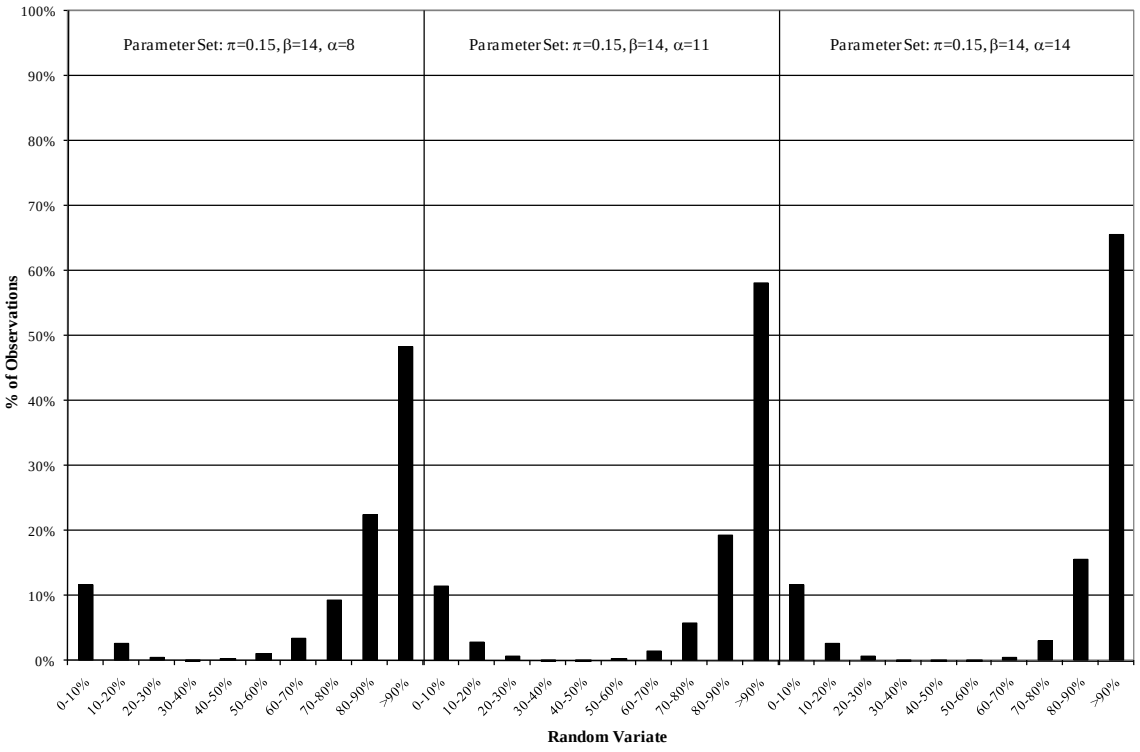

Parameter Set: π=0.15, β=2, α=8
Parameter Set: π=0.15, β=2, α=11
Parameter Set: π=0.15, β=2, α=14
Parameter Set: π=0.15, β=3, α=2
Parameter Set: π=0.15, β=3, α=3
Parameter Set: π=0.15, β=3, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

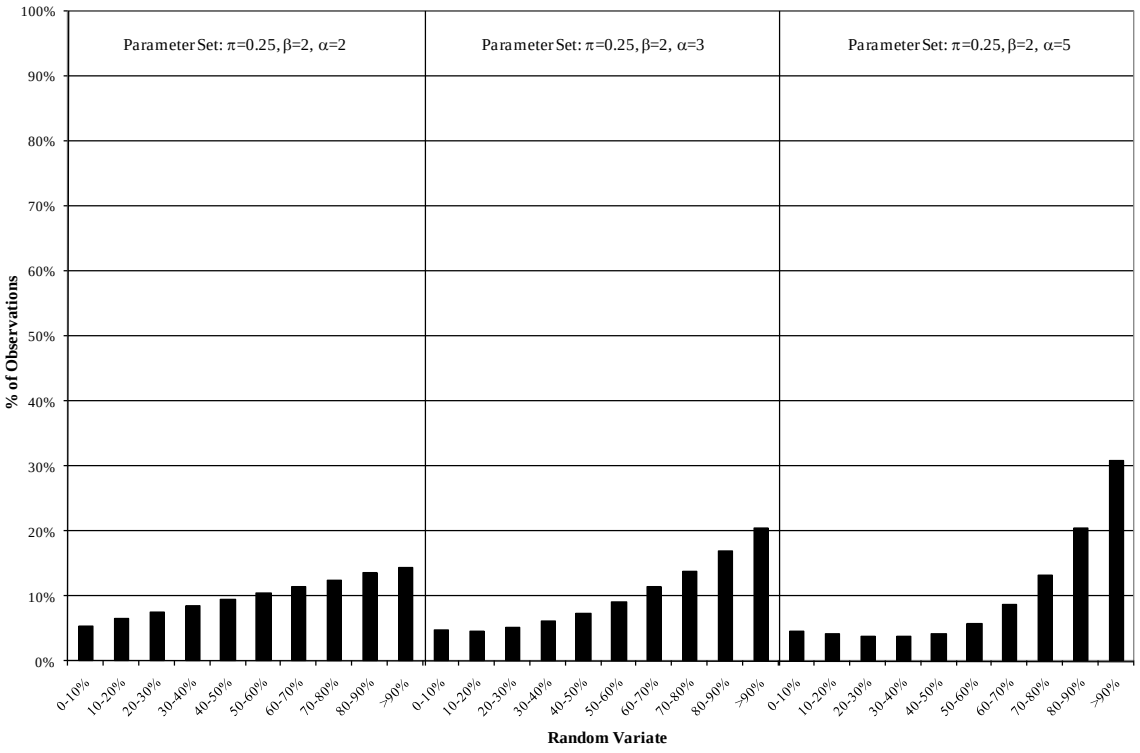

Parameter Set: π=0.15, β=3, α=8
Parameter Set: π=0.15, β=3, α=11
Parameter Set: π=0.15, β=3, α=14
Parameter Set: π=0.15, β=5, α=2
Parameter Set: π=0.15, β=5, α=3
Parameter Set: π=0.15, β=5, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

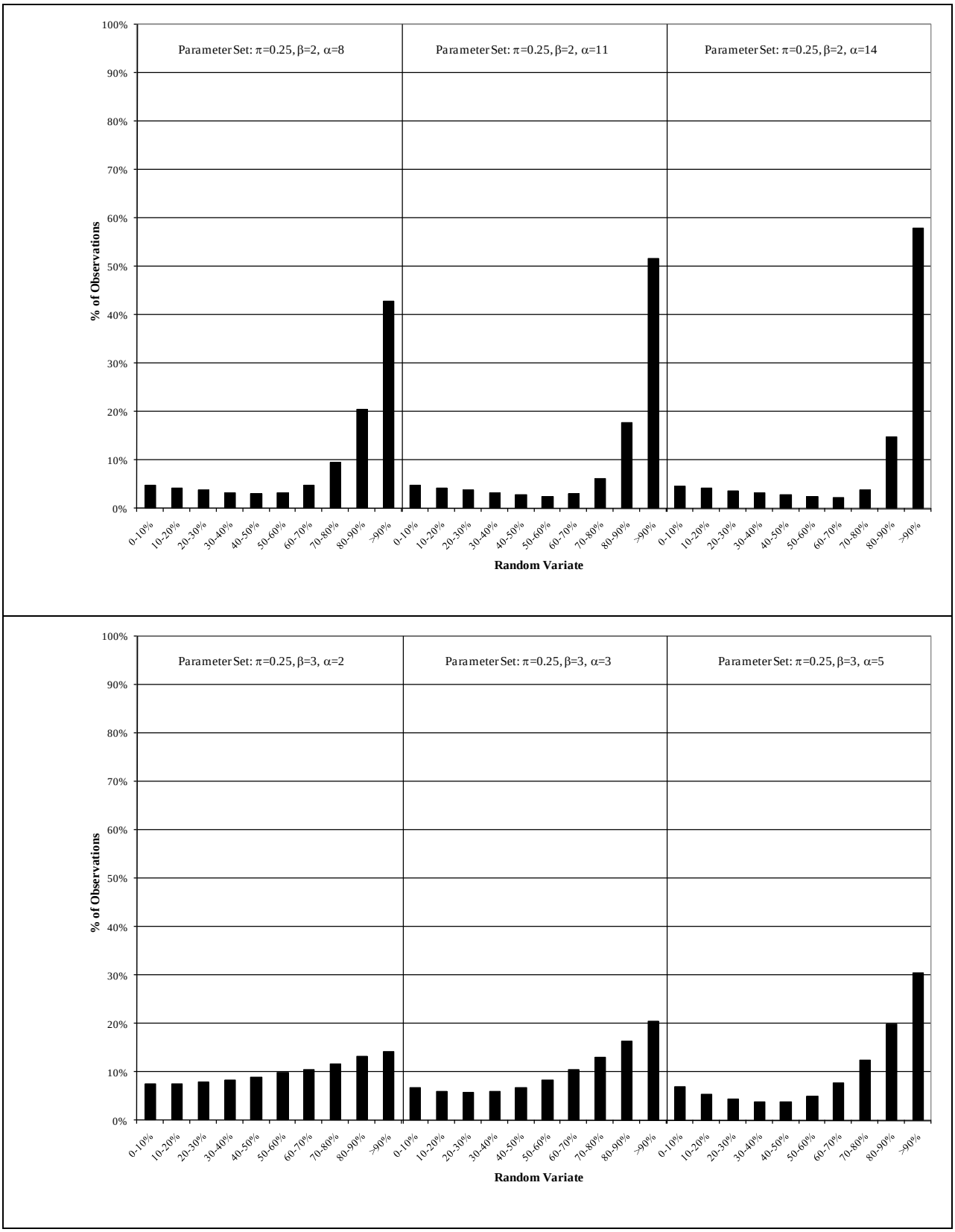

Parameter Set: π=0.15, β=5, α=8
Parameter Set: π=0.15, β=5, α=11
Parameter Set: π=0.15, β=5, α=14
Parameter Set: π=0.15, β=8, α=2
Parameter Set: π=0.15, β=8, α=3
Parameter Set: π=0.15, β=8, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

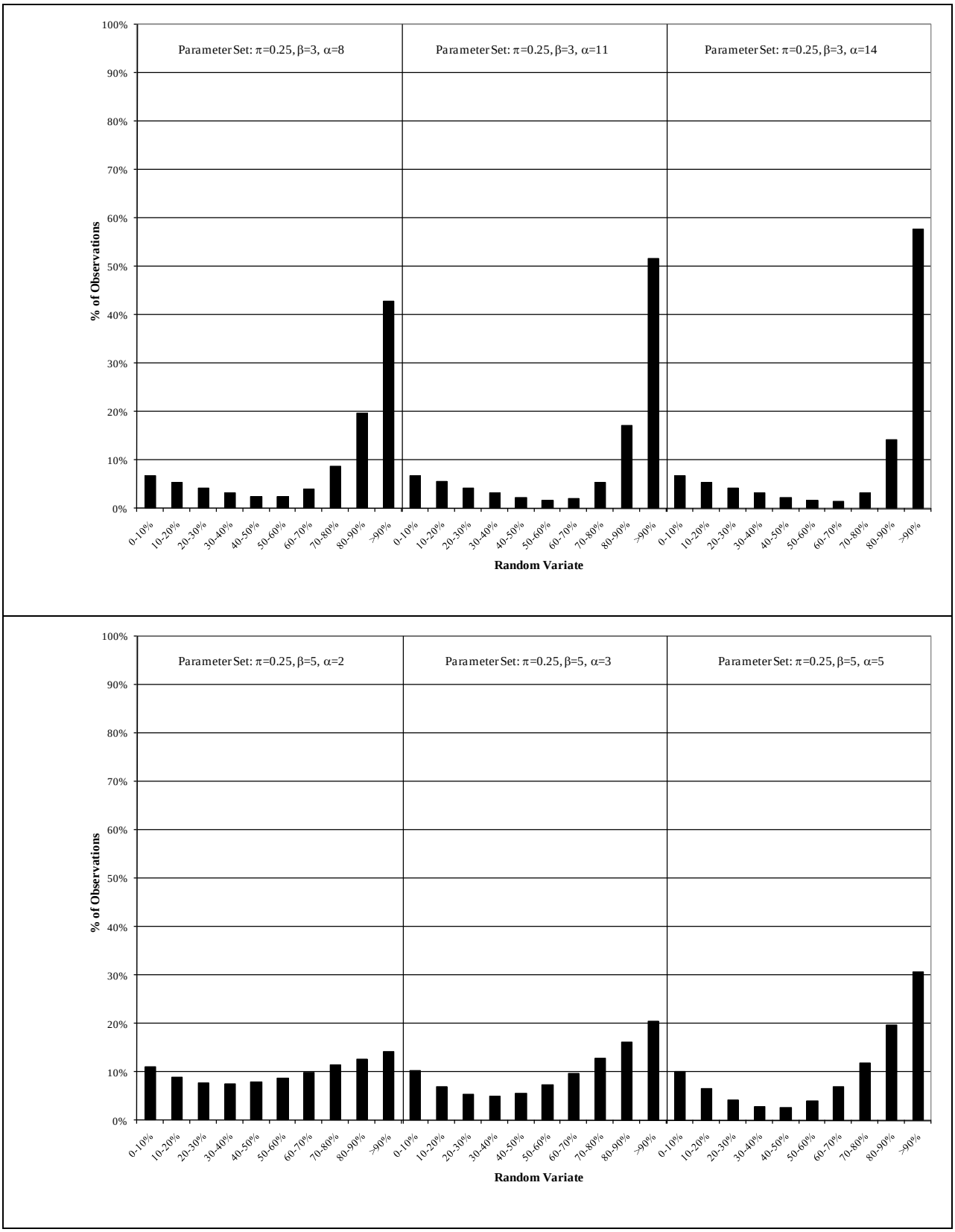

Parameter Set: π=0.15, β=8, α=8
Parameter Set: π=0.15, β=8, α=11
Parameter Set: π=0.15, β=8, α=14
Parameter Set: π=0.15, β=11, α=2
Parameter Set: π=0.15, β=11, α=3
Parameter Set: π=0.15, β=11, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

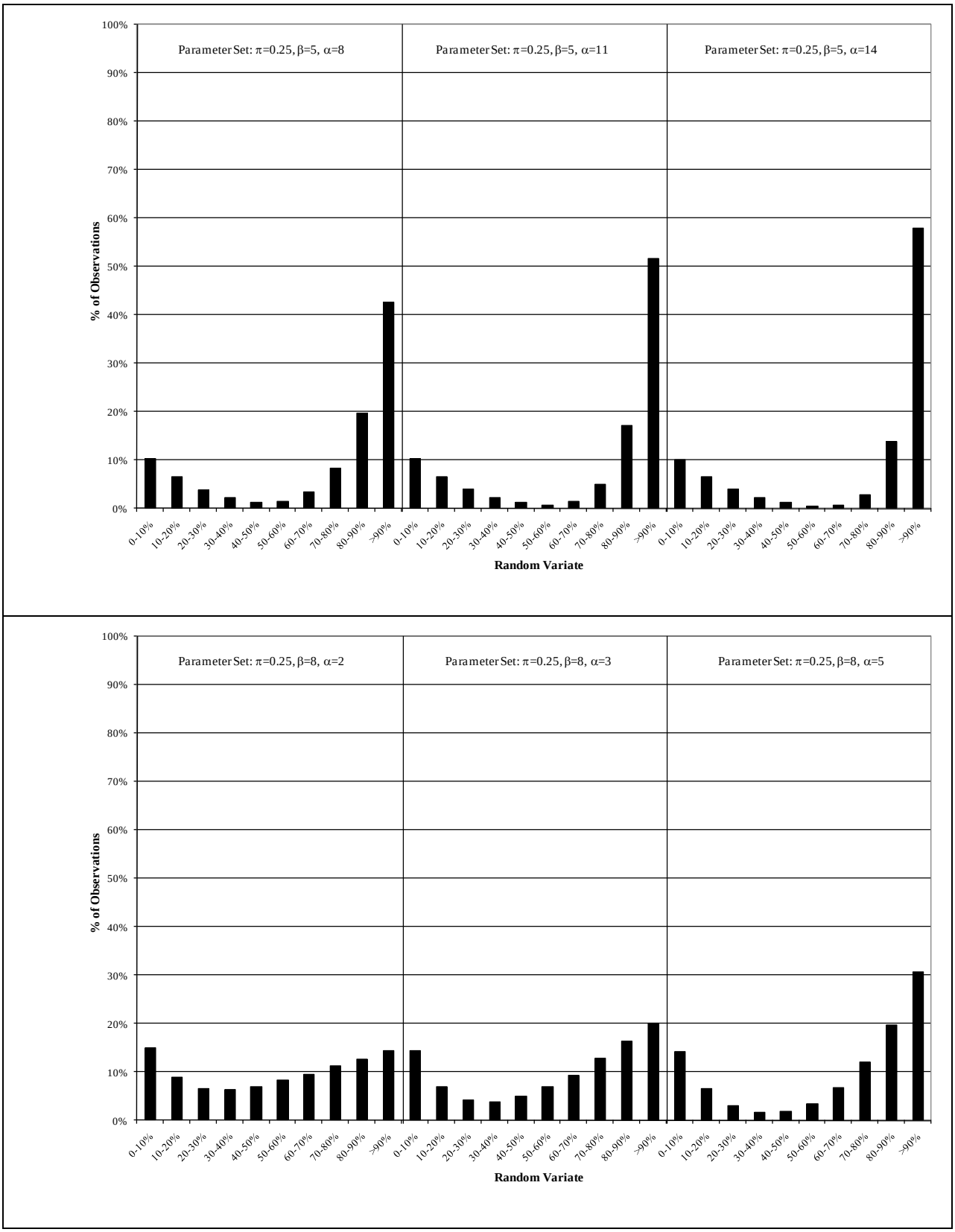

Parameter Set: π=0.15, β=11, α=8
Parameter Set: π=0.15, β=11, α=11
Parameter Set: π=0.15, β=11, α=14
Parameter Set: π=0.15, β=14, α=2
Parameter Set: π=0.15, β=14, α=3
Parameter Set: π=0.15, β=14, α=5
% of Observations
Random Variate
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

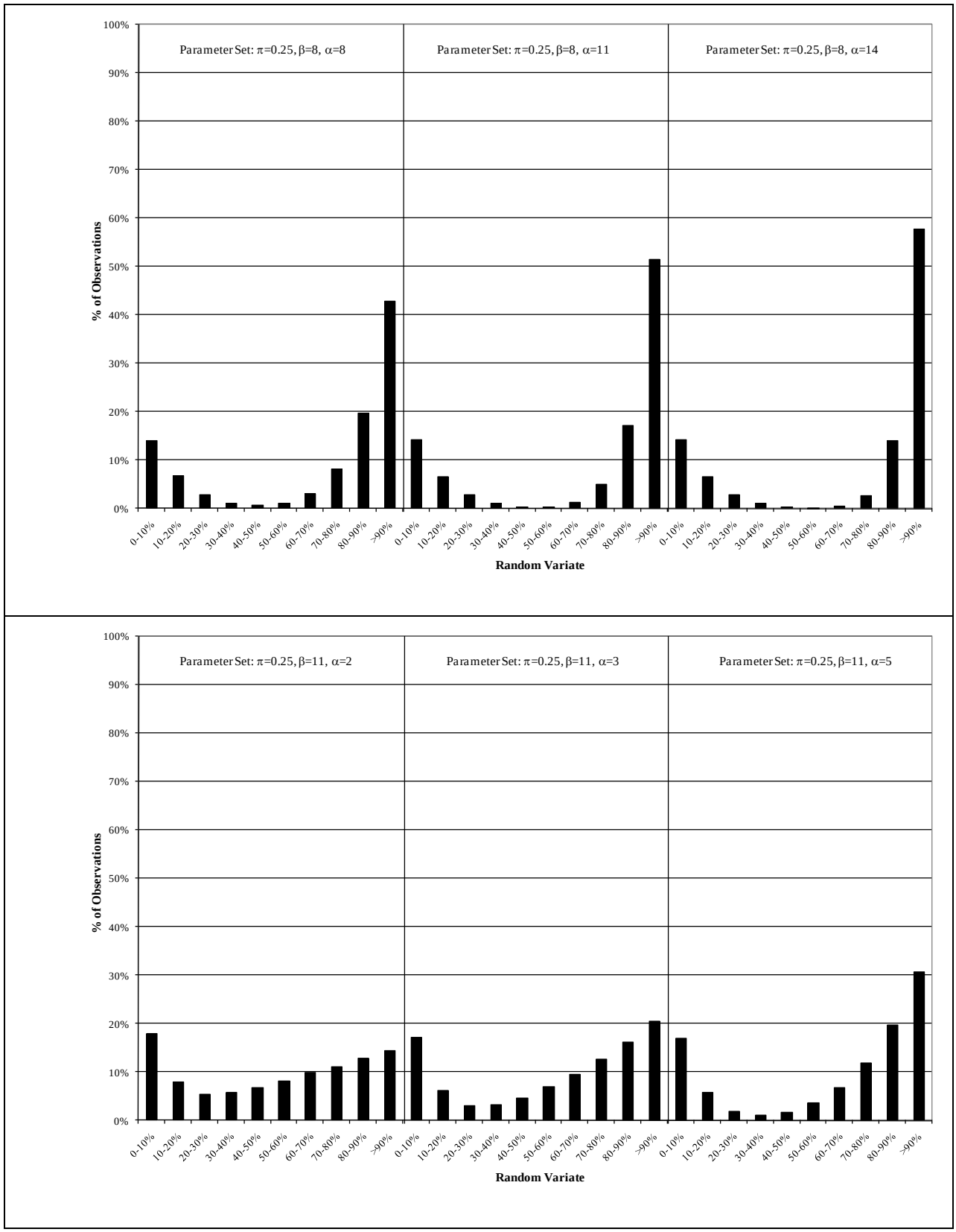

Parameter Set: π=0.15, β=14, α=8
Parameter Set: π=0.15, β=14, α=11
Parameter Set: π=0.15, β=14, α=14
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

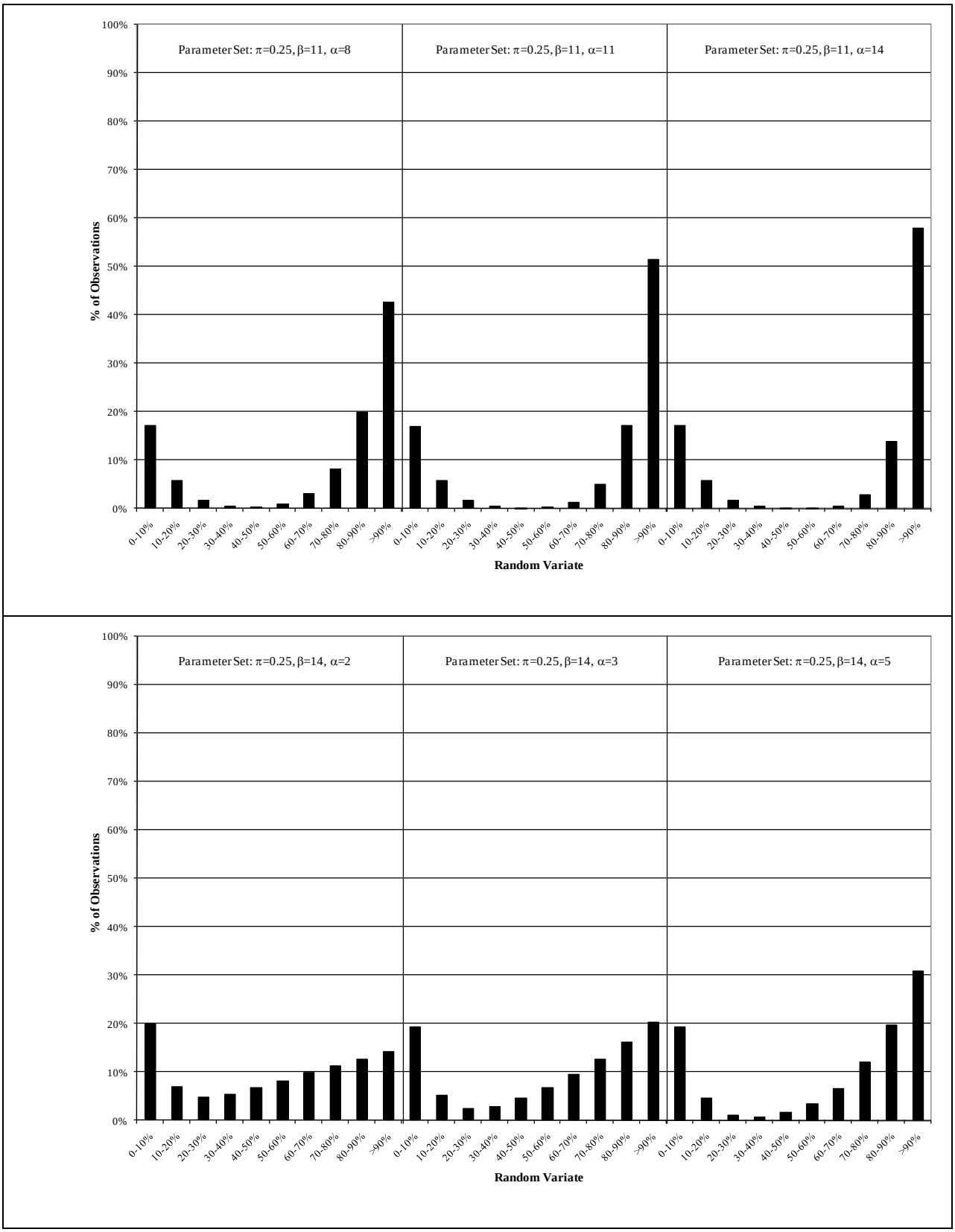

Parameter Set: π=0.25, β=2, α=2
Parameter Set: π=0.25, β=2, α=3
Parameter Set: π=0.25, β=2, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%

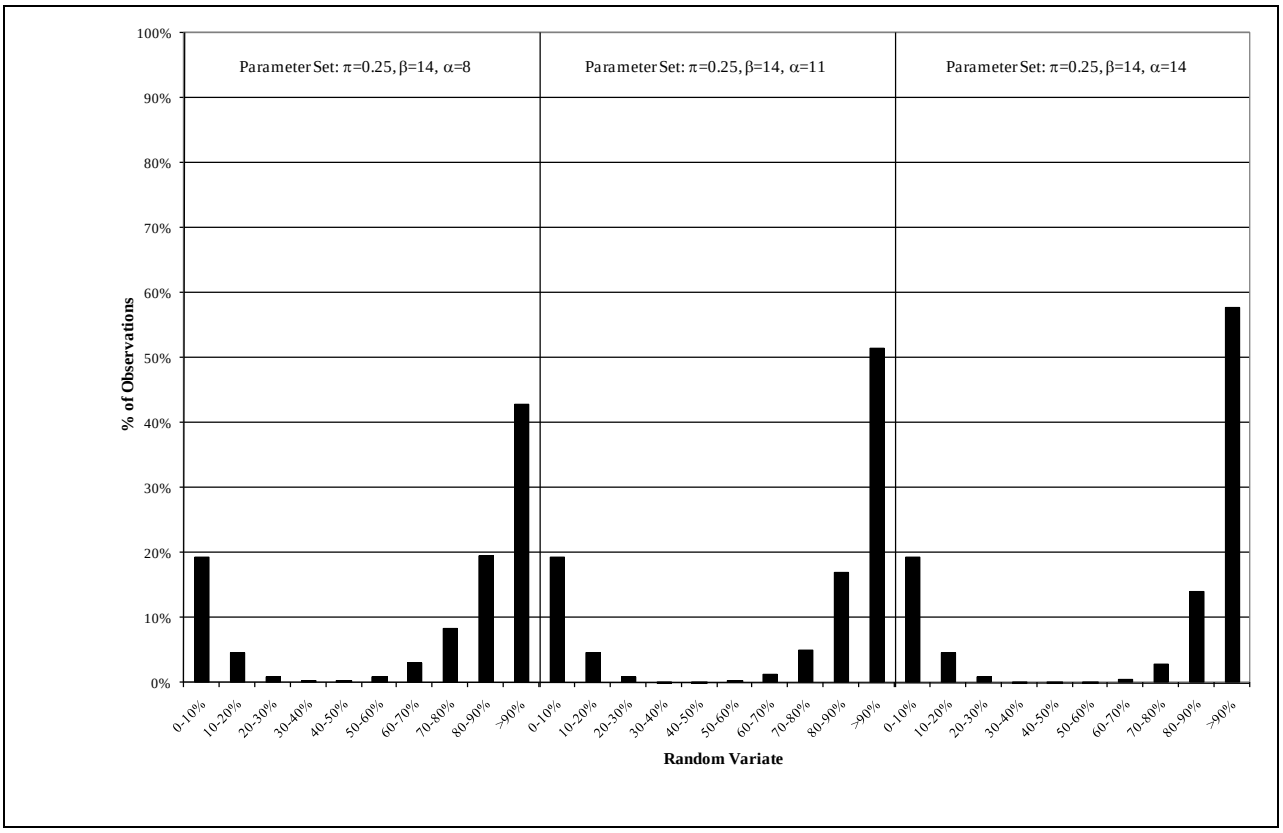

Parameter Set: π=0.25, β=2, α=8
Parameter Set: π=0.25, β=2, α=11
Parameter Set: π=0.25, β=2, α=14
Parameter Set: π=0.25, β=3, α=2
Parameter Set: π=0.25, β=3, α=3
Parameter Set: π=0.25, β=3, α=5
% of Observations
Random Variate
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%


Parameter Set: π=0.25, β=3, α=8
Parameter Set: π=0.25, β=3, α=11
Parameter Set: π=0.25, β=3, α=14
Parameter Set: π=0.25, β=5, α=2
Parameter Set: π=0.25, β=5, α=3
Parameter Set: π=0.25, β=5, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%


Parameter Set: π=0.25, β=5, α=8
Parameter Set: π=0.25, β=5, α=11
Parameter Set: π=0.25, β=5, α=14
Parameter Set: π=0.25, β=8, α=2
Parameter Set: π=0.25, β=8, α=3
Parameter Set: π=0.25, β=8, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%


Parameter Set: π=0.25, β=8, α=8
Parameter Set: π=0.25, β=8, α=11
Parameter Set: π=0.25, β=8, α=14
Parameter Set: π=0.25, β=11, α=2
Parameter Set: π=0.25, β=11, α=3
Parameter Set: π=0.25, β=11, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%


Parameter Set: π=0.25, β=11, α=8
Parameter Set: π=0.25, β=11, α=11
Parameter Set: π=0.25, β=11, α=14
Parameter Set: π=0.25, β=14, α=2
Parameter Set: π=0.25, β=14, α=3
Parameter Set: π=0.25, β=14, α=5
% of Observations
Random Variate
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%


Parameter Set: π=0.25, β=14, α=8
Parameter Set: π=0.25, β=14, α=11
Parameter Set: π=0.25, β=14, α=14
% of Observations
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
>90%
Random Variate

BIBLIOGRAPHY

Allport, F. H. (1934). The J-curve of the conforming hypothesis. *Journal of Social Psychology*, *5*, 141-183.

Aronson, B. S. (2006). Antibiotic-taking experiences of undergraduate college students. *Journal of the American Academy of Nurse Practitioners*, *18*, 591-598.

Bartlett, J. A. (2002). Addressing the Challenges of Adherence. *Journal of Aquired Immune Deficiency Syndromes*, *29*, s2-s10.

Bartlett, S. J., Lukk, P., Butz, A., Lampros-Klein, F., & Rand, C. S. (2002). Enhancing medication adherence among inner-city children with asthma: Results from pilot studies. *Journal of Asthma*, *39* (1), 47-54.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, *Series B 26*, 211-252.

Bramley, T. J., Gerbino, P. P., Nightengale, B. S., & Frech-Tamas, F. (2006). Relationship of Blood Pressure Control to Adherence with Antihypertensive Monotherapy in 13 Managed Care Organizations. *Journal of Managed Care Pharm*, *12* (3), 239-245.

Cassella, G., & Berger, R. L. (1990). *Statistical Inference.* Belmont, CA, USA: Duxbury Press.

Cramer, J. A., Scheyer, R. D., & Mattson, R. H. (1990). Compliance declines between clinic visits. *Archives of Internal Medicine*, *150*, 1509-1510.

Dodd, S. C. (1949). Notes on an Index of Conformity. *Science*, *110* (2853), 233-235.

Dolan, B. (2009, August 15). *Poor medication adherence costs $290 billion a year.* Retrieved October 28, 2009, from Mobile Health News Web site: http://mobihealthnews.com/3901/

Dunbar, J. (1990). Predictors of patient adherence: Patient characteristics. In J. K. Ockene, S. Schumaker, & E. Schron, *The Adoption and Maintenance of Behaviors for Optimal Health.* New York: Springer.

Dunbar-Jacob, J., Sereika, S., Burke, L. E., Starz, T., Rohay, J. M., & Kwoh, C. K. (1995). Can Poor Adherence Be Improved? *16th Annual Scientific Sessions of the Society of Behavioral Medicine.* San Diego, CA.

Dunbar-Jacob, J., Sereika, S., Rohay, J. M., & Probstfield, J. (1994). J-Shaped compliance distribution: Revisited. *Society for Clinical Trials*, *15*, p. 120S.

Efron, B. (1982). *The Jackknife, Bootstrap and Other Resampling Plans.* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap.* Boca Raton, FL, USA: CRC Press LLC.

Esposito, L. (1995). The Effects of Medication Education on Adherence to Medication Regimens in an Elderly Population. *Journal of Advanced Nursing*, *21*, 935-943.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical Distributions.* New York, NY, USA: John Wiley & Sons, Inc.

Gallagher, E. J., Viscoli, C. M., & Horwitz, R. I. (1993). The relationship of treatment adherence to the risk of death after myocardial infarction in women. *The Journal of the American Medical Association*, *270*, 742-744.

Gautschi, W. (1997). The Computation of Special Functions by Linear Difference Equations. In S. Elaydi, I. Gyori, & G. Ladas, *Advances in Difference Equations: Proceedings of the Second International Conference on Difference Equations* (pp. 213-244). Amsterdam: Overseas Publishers Association.

Horwitz, R. I., Viscoli, C. M., Berkman, L., Donaldson, R. M., Horwitz, S. M., Murray, C. J., et al. (1990). Treatment adherence and risk of death after a myocardial infarction. *Lancet*, *336*, 542-545.

Ji, Y., Wu, C., Liu, P., Wang, J., & Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, *21* (9), 2118-2122.

Jonasson, G., Carlsen, K.-H., Sodal, A., Jonasson, C., & Mowinckel, P. (1999). Patient compliance in a clinical trial with inhaled budesonide in children with mild asthma. *European Respiratory Journal*, *14*, 150-154.

Lange, K. (1998). *Numerical Analysis for Statisticians.* New York, NY, USA: Springer-Verlag.

Matsui, D., Hermann, C., Klein, J., Berkovitch, M., Olivieri, N., & Koren, G. (1994). Critical Comparison of Novel and Exisiting Methods of Compliance Assessment During a Clinical Trial of an Oral Iron Chelator. *Journal of Clinical Pharmacology*, *34*, 944-949.

McDonald, H. P., Garg, A. X., & Haynes, R. B. (2002). Interventions to Enhance Patient Adherence to Medication Prescriptions. *Journal of the American Medical Association*, *288*, 2868-2879.

McKenney, J. M., Munroe, W. P., & Wright, J. T. (1992). Impact of an Electronic Medication Compliance Aid on Long-Term Blood Pressure Control. *Journal of Clinical Pharmacology*, *32*, 277-283.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference.* Newbury Park, CA, USA: Sage Publications, Inc.

Norrell, S. E. (1981). Monitoring compliance with pilocarpine therapy. *American Journal of Ophthalmology*, *92*, 727-731.

Olivieri, N. F., Matsui, D., Hermann, C., & Koren, G. (1991). Compliance Assessed by the Medication Event Monitoring System. *Archives of Disease in Childhood*, *66*, 1399-1402.

Pachter, L. M., & Weller, S. C. (1993). Acculturation and compliance with medical therapy. *The Journal of Developmental and Behavioral Pediatrics*, *14*, 163-168.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford: Oxford University Press.

Pizzo, P. A., Robichaud, K. J., Edwards, B. K., Schumaker, C., Kramer, B. B., & Johnson, A. (1983). Oral antibiotic prophylaxis in patients with cancer: A double-blind randomized placebo-controlled trial. *The Journal of Pediatrics*, *102*, 125-133.

Rohay, J. M., Dunbar-Jacob, J., Sereika, S., & Kwoh, K. (1996). The Impact of Method of Calculation of Electronically Monitored Adherence Data. *Controlled Clinical Trials*, *17*, 82S-83S.

Rosner, B. (1990). *Fundamentals of Biostatistics* (3rd Edition ed.). Boston, MA, USA: PWS-Kent Publishing Company.

SAS Institute Inc. (2002-2003). Version 9 of the SAS System for Windows. Cary, NC, USA.

Solomon, R. S. (1939). Further Theoretical Considerations of the J-Curve Hypothesis. *Sociometry*, *2* (2), 59-68.

The Coronary Drug Project Research Group. (1980). Influence of adherence to treatment and response of cholesterol on mortality in The Coronary Drug Project. *New England Journal of Medicine*, *303*, 1038-1041.

Zygmunt, A., Olfson, M., Boyer, C. A., & Mechanic, D. (2002). Interventions to Improve Medication Adherence in Schizophrenia. *American Journal of Psychiatry*, *159*, 1653-1664.